Національний технічний університет України
«Київський політехнічний інститут імені Ігоря Сікорського»
Міністерство освіти і науки України

Національний технічний університет України
«Київський політехнічний інститут імені Ігоря Сікорського»
Міністерство освіти і науки України

Кваліфікаційна наукова
праця на правах рукопису

ГУСЬКОВА ВІРА ГЕННАДІЇВНА

УДК 004.896:004.622:004.891

ДИСЕРТАЦІЯ

МЕТОДИ І МОДЕЛІ ІНТЕЛЕКТУАЛЬНОГО АНАЛІЗУ ДАНИХ ДЛЯ ОЦІНЮВАННЯ ФІНАНСОВИХ РИЗИКІВ

122 – «Комп’ютерні науки»
12 – «Інформаційні технології»

Подається на здобуття наукового ступеня доктора філософії

Дисертація містить результати власних досліджень. Використання ідей, результатів
і текстів інших авторів мають посилання на відповідне джерело
_____В. Г. Гуськова

Науковий керівник
Бідюк Петро Іванович, доктор технічних наук, професор.

Київ – 2020

АНОТАЦІЯ

Гуськова В. Г. Методи і моделі інтелектуального аналізу даних для оцінювання фінансових ризиків. − Кваліфікаційна наукова праця на правах рукопису.

Дисертація на здобуття наукового ступеня доктора філософії за спеціальністю 122 «Комп'ютерні науки» – Національний технічний університет України «Київський політехнічний інститут імені Ігоря Сікорського», Київ, 2020.

Економічний успіх країн забезпечується застосуванням наукових підходів при управлінні процесами у будь-якій галузі практичної діяльності. Використання сучасних методів математичного моделювання, зокрема, методів інтелектуального аналізу даних до розв'язання проблем прогнозування і моделювання на сьогодні є актуальним завданням. Ризики притаманні будь-якій фінансовій установі або підприємству, а тому належне управління ризиками є важливим аспектом. Керівництво банку або компанії має різний рівень контролю за ризиками, тож внаслідок цього деякими з них можна безпосередньо керувати, а деякі інші ризики знаходяться значною мірою поза контролем управлінців. Найкраще, що можна зробити для запобігання ризиків, – це передбачити їх, оцінити потенційний вплив та бути готовим до реалізації плану реагування на несприятливі події.

Також важливим є оцінювання станів в умовах невизначеностей та за відсутністю методик аналізу розвитку нелінійних нестаціонарних процесів. Даний підхід дозволяє розв'язати велику кількість складних завдань, які відносяться до різних видів людської діяльності. Перш за все це застосовується у завданнях, які характеризується великою кількістю чинників або мають при постанові фактори невизначеності. Фінансові ризики відносяться до специфічної сфери кредитно- та ринково-фінансових взаємовідношень банку, фізичних та юридичних осіб, кінцевою метою яких, з точки зору організації, є отримання прибутку. Такі відносини чутливі до впливу великої кількості різноманітних,

часом невизначених ризиків, які у повсякденній роботі фінансової установи мають бути мінімізовані. Основною метою є розробка та вдосконалення математичних методів інтелектуального аналізу даних, які базуються на регресійних моделях, нейронних мережах, мережі Байєса та деревах рішень та відрізняються попередньою обробкою та згладженими початковими даними, що веде до підвищення точності задач фінансових ризиків. Практична цінність роботи полягає у тому, що: розроблено методи і моделі інтелектуального аналізу даних для оцінювання фінансових ризиків. Всі результати роботи доведено до практичного інженерного рівня і впроваджено у навчальний процес інституту прикладного системного аналізу НТУУ «КПІ імені Ігоря Сікорського» та у фінансових організаціях з метою автоматизованого розв'язання задач моделювання, оцінювання і прогнозування втрат у фінансовій компанії як стохастичного багатовимірного процесу. Створено системну методику та систему підтримки прийняття рішень для моделювання, оцінювання і прогнозування ринкового та кредитного фінансових ризиків. Всі теоретичні і практичні результати дисертаційної роботи у повній мірі опубліковано у фахових вітчизняних та закордонних наукових виданнях, що входять до відповідного встановленого переліку, а також виконано їх належну апробацію на міжнародних наукових конференціях і семінарах.

У дисертаційній роботі проаналізовано основні фінансові ризики, управління якими є ключовим фактором, що визначає ефективність фінансової діяльності. Виконано аналіз діяльності банків та інших фінансових організацій, робота яких здійснюється під впливом невизначеностей зовнішнього середовища (ринку, економіки, політики тощо), великої кількості змінних, контрагентів, осіб, поведінка яких не завжди може бути передбачена з прийнятною точністю. Розглянуто можливість мінімізації фінансових ризиків на двох рівнях – на рівні кожної окремої позики та на рівні кредитного портфеля в цілому. В результаті виконання попереднього аналізу встановлено, що найпоширенішими методами оцінювання ризиків для поставленої задачі є лінійна і логістична регресії, дерева класифікації, нейронні мережі, мережа

Байєса. Показано, що для підвищення ефективності прийняття об'єктивних рішень при аналізі кредитного та ринкового ризиків доцільно використовувати мережі Байєса та нечіткі нейронні мережі, які дають можливість враховувати невизначеності ймовірнісного та амплітудного типів. Ці підходи характеризуються швидкими алгоритмами навчання та нескладною інтерпретацією накопичених знань. Такі особливості обраних підходів роблять їх одними з найбільш перспективних і ефективних інструментів моделювання і оцінювання фінансових ризиків.

Розроблено метод оцінювання кредитоспроможності фізичних осіб та запропоновано підхід для вибору підмножини ознак, який характеризується оцінкою значущості результатів за сукупністю підходів та дозволяє підвищити відсоток вірних класифікацій на 3-4% що підтверджується його практичним застосуванням і отриманими результатами експериментальних досліджень. Показано, що процес аналізу кредитоспроможності позичальників кредитів включає в себе розробку методів та критеріїв аналізу процесу кредитування; оцінку потенційного клієнта, а також супровід позики після видачі кредиту. Визначено, що банкам та організаціям для оптимізації кредитного процесу необхідно знаходити компроміс між якістю та ефективністю всього процесу кредитування, а оцінка кредитного ризику здійснюється за допомогою оцінки кредитоспроможності позичальника. Для обчислення нечіткого висновку побудовано нечітку базу знань та використано ННМ в яких результат отримуються на основі нечітких логічних висновків, а параметри функцій належності налаштовуються за допомогою алгоритмів нейронних мереж. Пріоритетною являється задача мінімізації ризику, через що відразу відкидаються ненадійні позичальники, а для всіх інших розв’язується задача максимізації доходу портфеля позик.

Удосконалено метод оцінювання кредитоспроможності позичальника з використанням Байєсівської мережі з урахуванням підвищення якості ймовірнісної моделі та зменшенням величини кредитного ризику.

Розроблено метод оцінювання ринкового ризику на основі інтегрованого застосування ймовірнісної, оптимальної та цифрової фільтрації і регресійної

моделі, який відрізняється високою якістю попередньої обробки даних і забезпечує підвищення якості оцінок прогнозів. Запропоновано підхід із використанням попередньої обробки даних у вигляді фільтрації із застосуванням ймовірнісного, оптимального та цифрового фільтрів та регресійної моделі, за результатами якого забезпечено підвищення якості оцінок прогнозів. Виконано експериментальні дослідження із використанням попередньої обробки даних, які показали, що даний підхід в середньому в 2-3 рази ефективніший ніж робота з даними без використання фільтрації. Показано, що процес попередньої обробки даних за допомогою фільтрації є дуже важливим етапом аналізу даних. Застосування методів на цьому етапі, дає можливість значно покращити результати досліджень. Інколи відсутність методів попередньої обробки ставить під загрозу всі подальші кроки по обробці даних. Це може призводити до низької якості результатів, наприклад, оцінки прогнозів характеризуються великими похибками. Проаналізовано та виконано адаптивне прогнозування нелінійних нестаціонарних процесів, яке є також однією з ключових задач сучасності, у зв'язку з тим, що більшість процесів в економіці, фінансах, екології та технологічних процесах дуже швидко змінюються та не мають єдиного підходу.

Запропоновано підхід із використанням адаптації математичної моделі фінансового процесу до початкових даних ринкових ризиків із застосуванням триконтурної процедури адаптації та побудови моделі нелінійного процесу у вигляді лінійної та нелінійної компонент.

Ефективність усіх розроблених систем керування підтверджена результатами проведених імітаційних моделювань.

Ключові слова: фільтрація, нелінійні процеси, регресійні моделі, лінійні та нелінійні компоненти, триконтурна адаптація, адаптивна Байєсівська мережа, нейронна мережа, прогноз.

ABSTRACT

Huskova V.H. Data mining methods and models for evaluating financial risks.− Qualifying scientific work on the rights of the manuscript.

Dissertation for obtaining a scientific degree of a Ph.D. degree in specialty 122 «Computer Science». - National Technical University of Ukraine " Igor Sikorsky Kyiv Polytechnic Institute", Ministry of Education and Science of Ukraine, Kyiv, 2020.

Economic growth of countries has provided by scientific approaches to process management in any field of practice. Using of modern methods of mathematical modeling, such as methods of data mining to solve problems with forecasting and modeling today is one of the relevant tasks. The risks are always by any institution or business that's why proper management is an important aspect. The management of a bank or company has different levels of risk control, so as a result, some of them can be directly managed, and some other risks are largely beyond the control of managers. The best thing to prevent risks is to anticipate them, assess the potential impact and be prepared to implement an adverse response plan.

It is also important to assess the states for conditions of uncertainty and in the absence of methods for analyzing the development of nonlinear non-stationary processes. This approach allows you to solve a large number of complex problems related to various human activities. First of all, it is used in tasks that are characterized by a large number of reasons or have uncertainty factors. Financial risks relate to a specific area of credit - and market-financial relations of the bank, individuals and legal entities, the ultimate goal of which, from the point of view of the organization, is to make a profit. Such relationships are sensitive to the impact of a large number of different risks, which should be minimized in the daily work of a financial institution.

The main aim of the thesis is to develop and improve mathematical methods of data mining based on regression models, neural networks, Bayesian networks and decision trees and characterized by pre-processing and smoothed initial data, which leads to increased accuracy of financial risks.

In the work the following scientific results were obtained: The method of market risk assessment based on the integrated application of probabilistic, optimal and digital filtering and regression model has been developed, which is characterized by high quality data pre-processing and improves the quality of forecast estimates. An adaptation method of the mathematical model of the financial process to the data has been developed, which differs in the applying of the three-loop adaptation procedure and provides construction of the nonlinear process model in the form of linear and nonlinear components. Method of enhancing the quality of forecasting borrowers' solvency has been improved, which differs with the combined approach to the selection of repressors and the use of alternative forecasting methods, which provides optimization of weighting coefficient estimates for individual forecasts. The method of assessing the borrowers' solvency using an adaptive Bayesian network has been improved, which is characterized by increased adequacy of the probabilistic model and provides a reduction in the amount of credit risk.

The practical significance of the thesis results is in the developed methods and models of data analysis for financial risk assessment. The results of the thesis fulfillment were used in educational process of the MMSA Department of Institute for Applied System Analysis (IASA) at the National Technical University of Ukraine "Igor Sikorsky Kyiv Polytechnic Institute", and in the financial organizations in order to automatically solve the problems of modeling, estimating and forecasting losses in a financial company as a stochastic multidimensional process. A systematic methodology and decision support system for modeling, assessing, forecasting market and credit financial risks has been created. All theoretical and practical results of the dissertation published in professional domestic and foreign scientific journals included in the relevant list, as well as their proper approbation at international scientific conferences and seminars.

In the work analyzed the main financial risks, the management of which is a key factor that determines the effectiveness of financial activities. Implemented an analysis of the activities of banks and other financial organizations, whose work is carried out under the influence of environmental uncertainties (market, economy, politics, etc.), a large

number of variables, counterparties, individuals whose behavior can not always be predicted with acceptable accuracy. The possibility of minimizing financial risks at two levels has been considered - at the level of each individual loan and at the level of the loan portfolio as a whole.

As a result of preliminary analysis, it was found that the most common methods of risk assessment for the task are linear and logistic regressions, classification trees, neural networks, Bayesian network. It is shown that to increase the efficiency of making objective decisions in the analysis of credit and market risks, it is advisable to use Bayesian networks and fuzzy neural networks, which make it possible to take into account uncertainties of probabilistic and amplitude types. These approaches are characterized by fast learning algorithms and simple interpretation of accumulated knowledge. Such features of the chosen approaches make them one of the most promising and effective tools for modeling and assessing financial risks.

In the dissertation work the comparative analysis of efficiency for using statistical methods, logit and probit models, Bayesian networks, decision trees, the neural network with the indistinct logical conclusion of Mamdani and Sugeno which showed the efficiency of fuzzy neural networks credit analysis for individual borrowers. The estimation method of solvency for individual borrowers has been developed. An approach is proposed for selecting a subset of features, which is characterized by assessing the significance of the results on a set of approaches and allows increase substantially the level of correct classifications by 4-5%, which is confirmed by its practical use and the results of experimental studies.

It is shown that the process of credit analysis of borrowers includes the development of methods and criteria for analyzing the lending process; assessment of a potential client, as well as support of the loan after the credit. It is determined that banks and organizations need to find a compromise between the quality and efficiency of the entire lending process to optimize the credit process, and credit risk is assessed by assessing the borrower's creditworthiness. To calculate the fuzzy inference, a fuzzy knowledge base is constructed and FNN are used in which the result is obtained on the basis of fuzzy logical inferences and the parameters of membership function are configured using neural network algorithms.

The priority task is to minimize the risk, which immediately rejects unreliable borrowers, and for all others, the task of maximizing the income of the loan portfolio.

The improved method was proposed for assessing the solvency of a credit borrower using the Bayesian network, and taking into account improving the quality of the probabilistic model and reducing the amount of credit risk. The method of market risk assessment based on the integrated application of probabilistic, optimal, and digital filtering and regression model has been developed, which is characterized by the high quality of data pre-processing, and improved quality of the forecast estimates.

An approach using pre-processing of data in the form of filtering with the use of probabilistic, optimal, and digital filters and a regression model is proposed, as a result of which the quality of forecast estimates is improved. An approach using pre-processing of data in the form of filtering using probabilistic, optimal and digital filters and a regression model is proposed, as a result of which the quality of forecast estimates is improved.

Experimental studies were performed using pre-processing of data, which showed that this approach is on average 2-3 times more efficient than working with data without the use of filtering. It is shown that the process of pre-processing data by filtering is a very important step in data analysis. The use of methods at this stage makes it possible to significantly improve the results of research. Sometimes the lack of pre-processing methods jeopardizes all further steps in data processing. This can lead to poor quality results, for example, estimates of forecasts are characterized by large errors. Adaptive forecasting of nonlinear non-stationary processes is analyzed and performed, which is also one of the key tasks of today, due to the fact that most processes in economics, finance, ecology and technological processes change very quickly and do not have a single approach.

The efficiency of all the developed control systems is confirmed by the results of the simulations.

Keywords: filtering, forecast, nonlinear processes, regression models, linear and nonlinear components, tricycle adaptation, adaptive Bayesian network, neural network.

Список публікацій здобувача за темою дисертації:

Статті у періодичних наукових виданнях держав, які входять до

Організації економічного співробітництва та розвитку та/або Європейського Союзу

1. Zaychenko Yu.: Recognition of Objects on Optical Images in Medical Diagnostics Using Fuzzy Neural Network Neffclass / Yu. Zaychenko, V. Huskova // International Journal "Information Models and Analyses". 2016. – №5. pp. 13-22. *(Болгарія). (запропоновано підхід для розпізнавання предметів на медичних зображеннях у медичній діагностиці).*

Статті у наукових фахових виданнях України, які входять до міжнародних наукометричних баз даних

2. Tymoshchuk O. L. A combined approach to modeling nonstationary heteroscedastic processes / O. L. Tymoshchuk, V. H. Huskova, P. I. Bidyuk. // Radio Electronics, Computer Science, Control. – 2019. – №2. – С. 80–89. *(включено до наукометричної бази Web of Science). (запропоновано модифіковану методологію моделювання нелінійних нестаціонарних процесів та схему адаптації для побудови моделей).*

Статті у наукових фахових виданнях України

3. Бідюк П. І. Аналіз кредитоспроможності за допомогою методів інтелектуального аналізу даних / П. І. Бідюк, В. Г. Гуськова // Електронне моделювання. - 2019. - Т. 41, № 2. - С. 111-120. *(досліджено підхід до мінімізації ризику платоспроможності позичальника для банківської системи та інших фінансових компаній).*

4. Гуськова В. Г. Оцінювання кредитоспроможності позичальників кредитів методами інтелектуального аналізу даних / В. Г. Гуськова, П. І. Бідюк. // Міжнародній науково-технічний журнал «Системні дослідження та інформаційні технологі. – 2019. – № 2. – С. 31–48. (*проаналізовано основні методи математичного моделювання і оцінювання кредитних ризиків, запропоновано математичні моделі для аналізу кредитних ризиків індивідуальних позичальників на основі альтернативних методів).*

5. Гуськова В. Г. Розробка сценарного підходу на основі моделей інтелектуального аналізу даних / В. Г. Гуськова, П. І. Бідюк // Наукові праці

Донецького національного технічного університету. Серія : Інформатика, кібернетика та обчислювальна техніка. – 2016. – № 2. – С. 158-164. (*запропоновано підхід на основі логістичної регресії для аналізу кредитоспроможності позичальників з використанням пакета RStudio із аналізом та застосуванням фактичних вихідних даних для оцінювання кредитних ризиків у майбутньому).*

6. Гуськова В. Г. Система підтримки прийняття рішень для прогнозування фінансових процесів на основі принципів системного аналізу / В. Г. Гуськова, Данилов В.Я., П. І. Бідюк, О.Л. Жиров // Міжнародній науково-технічний журнал «Системні дослідження та інформаційні технологі. – 2019. – №1. – С.20–36. *(досліджено концепцію розв'язання задач адаптивного прогнозування на основі методології системного аналізу).*

7. Гуськова В. Г. Аналіз кредитоспроможності позичальників кредитів за допомогою логістичної регресії. / В. Г. Гуськова, П. І. Бідюк // Наукові праці Донецького національного технічного університету. Серія: Інформатика, кібернетика та обчислювальна техніка. – 2017. –№2. - С. 54-60. (*побудовано математичні моделі аналізу для коротко-, середньо- та довгострокових прогнозів; розроблено авторегресійні рівняння).*

Наукові праці, які засвідчують апробацію матеріалів дисертації

8. Бідюк П.І., Гуськова В.Г.: Застосування нечітких правил регресійного аналізу до фінансових даних, Institute for Modelling in Energy Engineering, NASc of Ukraine, NASc of Ukraine, September 12-14, 2018, Kyiv, Ukraine.

9. Huskova V.H., Bidyuk P.I.: Estimating financial risk using systemic approach, Проблеми інформатизації, тези доповідей шостої міжнародної науково-технічної конференції, 14 – 16 листопада 2018 року, Черкаси – Баку – Бельсько-Бяла – Харків.

10. Bidyuk P., Huskova V., Terentiev O.: Client solvency estimation using intellectual data analysis approach, The VIth International Conference «Advanced Information Systems and Technologies, AIST 2018», 16-18 May 2018, Sumy, Ukraine.

11. Zaychenko Yu., Huskova V.: Application of fuzzy neural network nefclass

for recognition of medical images in diagnostics, System Analysis and Information Technologies; 18-th International Conference SAIT, 30 May - 02 June 2016, Kyiv, Ukraine.

12. Huskova V., Bidyuk P.: A Combined Approach to Modeling Heteroscedastic Processes and Financial Risk Estimation, Всеукраїнська науково-практична конференція комп'ютерна інженерія і кібербезпека: досягнення та інновації 27–29 листопада 2018 року, Кропивницький, Україна.

13. Гуськова В.Г., Бідюк П.І.: Побудова сценаріїв із використанням байєсівських методів, міжнародна науково-технічна конференція "моделювання і комп'ютерна графіка", 18–24 вересня, 2017 року, Покровськ, Україна.

ЗМІСТ

ПЕРЕЛІК УМОВНИХ ПОЗНАЧЕНЬ

БМ – байєсівська мережа

VaR – Value-at-Risk

RWA – зважені з урахуванням ризику активи

LASSO – least absolute shrinkage and selection operator

RFE – recursive feature elimination

SVM – support vector machine

ФН – функція належності

ММ – математична модель

ННМ – нечіткій нейронній мережі

TSK – Takagi, Sugeno, Kang'a

ANFIS – adaptive network-based fuzzy inference system

МНК – метод найменших квадратів

АР – авторегресія

АРУГ – авторегресійна умовна гетероскедастичність

МКМЛ – метод Монте-Карло для марковських ланцюгів

ASIR – auxiliary sampling importance resampling filter

RPF – regularized particle filter

ВСТУП

Актуальність роботи. Побудова адекватних математичних моделей і розробка ефективних методів розв’язання задач моделювання і прогнозування фінансових процесів та оцінювання ризиків можливих втрат – одне із ключових завдань поточного етапу розвитку методів інтелектуального аналізу даних, прикладної статистики і методів аналізу часових рядів. На сьогодні прогрес у напрямі підвищення якості та розширення можливостей сучасних систем математичного моделювання, оптимального оцінювання станів, прогнозування розвитку стохастичних процесів в економіці, фінансах, технічних системах і технологіях, а також оцінювання фінансових ризиків різних типів неможливий без застосування сучасних методів системного аналізу та інформаційних технологій, коректної обробки статистичних даних і врахування множини наявних невизначеностей структурного, статистичного і параметричного характеру.

Досвід розвинених країн свідчить про те, що економічний успіх будь-якої країни неможливий без застосування наукового підходу на всіх рівнях управління процесами різної природи у будь-якій галузі практичної діяльності. Актуальною задачею є сьогодні застосування сучасних методів математичного моделювання, особливо методів інтелектуального аналізу даних, до розв’язання практичної задачі моделювання і прогнозування розвитку фінансових процесів та супутніх ризиків, оцінювання їх станів в умовах наявності випадкових впливів і невизначеностей, а також за відсутності ефективних методик поглибленого аналізу даних стосовно подальшого розвитку нелінійних нестаціонарних процесів. Завдяки цьому при розв’язанні цілого ряду складних задач у різних видах людської діяльності вдається досягнути результатів, недосяжних в інший спосіб. Особливо це стосується тих задач, при розв’язанні яких потрібно враховувати багато різноманітних чинників, або у постановці яких має місце невизначеність того чи іншого типу. Фінансові ризики відносяться до специфічної сфери кредитно- та ринково-фінансових взаємовідношень банку,

фізичних та юридичних осіб, кінцева мета яких, з точки зору організації, є отримання прибутку. Указані відносини піддаються великій кількості різноманітних, часом невизначених, ризиків, які у повсякденній роботі банку мають бути мінімізовані.

Мета та задачі дослідження. Метою роботи дисертаційного дослідження є підвищення якості оцінювання і прогнозування фінансових ризиків шляхом розробки та адаптації моделей і методів інтелектуального аналізу даних.

Для досягнення поставленої мети необхідно розв’язати такі задачі:

- проаналізувати наявні методи інтелектуального аналізу даних та їх недоліки стосовно аналізу фінансових ризиків, запропонувати шляхи їх подальшого удосконалення стосовно розв’язання поставлених задач;
- адаптувати та удосконалити метод оцінювання кредитоспроможності позичальників з використанням байєсівської мережі з метою забезпечення зменшення величини кредитного ризику;
- адаптувати та удосконалити метод підвищення якості оцінювання та прогнозування кредитоспроможності позичальників шляхом оптимізації вагових коефіцієнтів оцінок окремих прогнозів;
- розробити метод оцінювання ринкового ризику на основі інтегрованого застосування ймовірнісної та регресійної моделей для забезпечення підвищення якості оцінок прогнозів;
- розробити метод адаптації математичної моделі фінансового процесу до даних з метою забезпечення побудови адекватної моделі нелінійного процесу у вигляді лінійної та нелінійної компонент.

Об’єкт дослідження: фінансові процеси і ризики – структурування, моделювання та прогнозування можливих втрат, виявлення закономірностей розвитку цих процесів та взаємозв’язків між ними.

Предмет дослідження: математичні методи і моделі аналізу фінансових даних на основі регресійного та ймовірнісно-статистичного підходів(мережі Байєса, методи фільтрації даних, дерева рішень, нечітка логіка, регресійний аналіз).

Методи дослідження: теорія ймовірностей, математична статистика, теорія прийняття рішень, теорія байєсових статистичних рішень, нейронні мережі та нейронечіткі моделі, регресійний аналіз та дерева рішень.

Наукова новизна результатів, отриманих автором, полягає у такому:

- **вперше** розроблено метод оцінювання ринкового ризику на основі інтегрованого застосування ймовірнісної, оптимальної та цифрової фільтрації і регресійної моделі, який відрізняється високою якістю попередньої обробки даних і забезпечує підвищення якості оцінок прогнозів.
- **вперше** розроблено метод адаптації математичної моделі фінансового процесу до даних, який відрізняється застосуванням триконтурної процедури адаптації і забезпечує побудову моделі нелінійного процесу у вигляді лінійної та нелінійної компонент.
- **удосконалено** метод підвищення якості прогнозування кредитоспроможності позичальників який відрізняється комбінованим підходом до вибору регресорів та використанням альтернативних методів прогнозування, що забезпечує оптимізацію вагових коефіцієнтів оцінок окремих прогнозів.
- **удосконалено** метод оцінювання кредитоспроможності позичальників з використанням адаптивної байєсівської мережі, який відрізняється підвищеною адекватністю ймовірнісної моделі і забезпечує зменшення величини кредитного ризику.

Обґрунтованість і достовірність наукових результатів забезпечується коректним застосуванням математичного апарату та відповідних обчислювальних експериментів, методів інтелектуального аналізу даних, прикладної статистики і методів аналізу часових рядів, а також процедурами контролю точності проміжних та остаточних результатів, що оцінювались на незалежних тестових вибірках і шляхом порівняння з даними офіційної статистики.

Практичне значення отриманих результатів полягає у такому:

Всі результати роботи доведено до практичного інженерного рівня і впроваджено у навчальний процес інституту прикладного системного аналізу

НТУУ «КПІ імені Ігоря Сікорського» та у компанії «ВІСОФТЮ» з метою автоматизованого розв'язання задач моделювання, оцінювання і прогнозування втрат у фінансовій компанії як стохастичного багатовимірного процесу. Створено системну методику та систему підтримки прийняття рішень для моделювання, оцінювання і прогнозування ринкового та кредитного фінансових ризиків. Всі теоретичні і практичні результати дисертаційної роботи у повній мірі опубліковано у фахових вітчизняних та закордонних наукових виданнях, що входять до відповідного встановленого переліку, а також виконано їх належну апробацію на міжнародних наукових конференціях і семінарах.

Особистий внесок здобувача. Всі основні результати дисертаційної роботи отримані автором особисто. Всі 7 фахових публікацій написано у співавторстві, а здобувачеві належать такі результати. В роботі [1] здобувачем запропоновано підхід до мінімізації ризику платоспроможності позичальника для банківської системи та інших фінансових компаній, які надають кредити своїм клієнтам, а також виконано оцінювання кредитоспроможності клієнтів з використанням логістичної регресії, методів на основі нечіткої логіки, нейронної мережі із зворотнім поширенням похибки і дерев рішень; надано результати оцінювання кредитоспроможності позичальників і проведено аналіз оцінювання стану клієнтів. В роботі [2] здобувачем проаналізовано основні методи математичного моделювання і оцінювання кредитних ризиків, запропоновано математичні моделі для аналізу кредитних ризиків індивідуальних позичальників на основі альтернативних методів, розроблено математичні моделі для аналізу кредитних ризиків індивідуальних позичальників на основі дерев рішень, логістичної регресії, мереж Байєса та нечіткої логіки. В роботі [3] здобувачем запропоновано модифіковану методологію моделювання нелінійних нестаціонарних процесів, схему адаптації для побудови адекватних математичних моделей, запропоновані нові модельні структури. В роботі [4] здобувачем запропоновано підхід на основі логістичної регресії для аналізу кредитоспроможності позичальників з використанням пакета RStudio із аналізом та застосуванням фактичних вихідних даних для оцінювання кредитних ризиків

у майбутньому; виконано порівняльний аналіз отриманих моделей та статистичних критеріїв і наведено приклад роботи даного підходу з використанням фінансових даних. В публікації [5] здобувачем запропоновано концепцію розв'язання задач адаптивного прогнозування на основі методології системного аналізу, що ґрунтується на комплексному використанні методів попередньої обробки даних, математичного і статистичного моделювання, прогнозування та оптимального оцінювання станів досліджуваних процесів. В публікації [6] здобувачем побудовані математичні моделі аналізу для коротко-, середньо- та довгострокових прогнозів; розроблено авторегресійні рівняння, авторегресійні рівняння з ковзним середнім (де ковзне середнє побудовано за залишками авторегресійної моделі), авторегресійна модель з ковзним середнім із альтернативними коефіцієнтами при ковзному середньому. Проаналізовано підхід на основі нейронних мереж з радіально-базисними функціями та нечітким методом групового урахування аргументів. В публікації [7] здобувачем виконано побудову математичних моделей у формі авторегресії першого і третього порядку на основі статистичних даних для країн світу; результати обчислювальних експериментів стосовно визначення оцінок трикрокових прогнозів свідчать, що моделі забезпечують отримання високоякісних оцінок прогнозів.

Апробація результатів дисертації. Результати та основні положення роботи доповідалися та обговорювалися на: The Sixth International Scientific Conference "Simulation-2018" (Kyiv, Ukraine, 12-14.09.2018) [8]; Шостій міжнародній науково-технічній конференції "Проблеми інформатизації" (Черкаси – Баку – Бельсько-Бяла – Харків, 14-16.11.2018) [9]; The VI-th International Conference «Advanced Information Systems and Technologies, AIST 2018» (Sumy, Ukraine, 16-18.05.2018) [10]; System Analysis and Information Technologies; 18-th International Conference SAIT (Kyiv, Ukraine, 30.05-02.06.2016) [11]; Всеукраїнській науково-практичній конференції "Комп'ютерна інженерія і кібербезпека: досягнення та інновації" (Кропивницький, Україна, 27-29.11.2018) [12]; VII міжнародна науково-технічна конференція "Моделювання і

комп’ютерна графіка" (Покровськ, Україна, 18-24.09.2017) [13].

Публікації. За матеріалами дисертації опубліковано 13 робіт, з яких 7 – це статті у журналах і збірниках наукових праць, що входять до переліку фахових видань затверджених МОН України за спеціальністю дисертації або у періодичних виданнях іноземних держав (1 включена до міжнародної наукометричної бази Web of Science), та 6 публікацій у матеріалах конференцій (у тому числі, міжнародних).

Структура та обсяг дисертації. Дисертація складається із анотації, вступу, чотирьох розділів, висновків, списку використаних джерел, додатків. Робота містить 199 сторінок, у тому числі: 171 сторінок основного тексту, 18 рисунків, 25 таблиць, список використаних джерел із 76 найменувань на 7 сторінках.

РОЗДІЛ 1
АКТУАЛЬНІСТЬ ДОСЛІДЖЕННЯ І МЕТОДИ ОЦІНЮВАННЯ ФІНАНСОВИХ РИЗИКІВ

1.1 Актуальність дослідження фінансового ризику та обґрунтування вибору кредитного і ринкового ризиків

Під фінансовим ризиком розуміють ймовірність виникнення непередбачуваних фінансових наслідків у формі втрати або отримання доходу в ситуації невизначеності умов здійснення фінансової діяльності юридичною або фізичною особою [14]. На сьогоднішній день ризики притаманні будь-якій фінансовій установі або підприємству, а тому належне управління (менеджмент) ризиками є важливим аспектом ведення успішного бізнесу. Керівництво банку або компанії має різний рівень контролю ризиків, а тому внаслідок цього деякими з них можна безпосередньо керувати, а деякі інші ризики знаходяться значною мірою поза контролем управлінців. Найкраще, що можна зробити в умовах настання ризику, – це передбачити можливі ризики, оцінити їх потенційний вплив та бути готовим до реалізації плану реагування на несприятливі події. Існує багато способів класифікувати фінансові ризики [15]. Один із підходів до цього забезпечується шляхом поділу фінансового ризику на чотири широкі категорії: ринковий ризик, кредитний ризик, ризик ліквідності та операційний ризик [16].

Ринковий ризик передбачає зміну умов перебування на конкретному ринку, на якому організації конкурують за свої сегменти ринку. Одним із прикладів ринкового ризику є зростаюча тенденція споживачів виконувати банкові операції в Інтернеті. Такий аспект ринкового ризику спричинив значні виклики для традиційного роздрібного бізнесу. Організації, які змогли виконати необхідні дії стосовно адаптації до обслуговування, процвітали і спостерігали значне зростання доходу, в той час як компанії, які повільно адаптувались або зробили некоректний вибір у своїй реакції на мінливий ринок, відійшли в сторону.

Цей приклад також стосується іншого елементу ринкового ризику – ризику перевищення рівня функціонування компанії конкурентами. На все більш конкурентоспроможному світовому ринку, хоча часто із зменшенням норм прибутку, найбільш фінансово успішні компанії пропонують унікальну цінову пропозицію, яка дозволяє їм виділятися з множини інших та надає їм надійну ринкову ідентичність.

Кредитний ризик – це ризик, з яким підприємства стикаються внаслідок надання кредитів клієнтам. Він також може стосуватися власного кредитного ризику компанії з постачальниками. Бізнес приймає фінансовий ризик, коли забезпечує фінансування закупівель для своїх клієнтів, надаючи можливість того, що клієнт може замовчуватись в оплаті [17]. Компанія повинна вирішувати свої власні кредитні зобов'язання, забезпечуючи, щоб у неї завжди був достатній грошовий потік та своєчасно сплачувати свої кредиторські рахунки. В іншому випадку постачальники можуть або припинити надавати кредит компанії, або взагалі припинити ведення бізнесу з цією компанією.

Базельський комітет з банківського нагляду визначає кредитний ризик як можливість того, що позичальник або котрагент банку не виконає свої зобов'язання у погоджений термін [18]. Мета управління кредитним ризиком полягає в максимізації норми прибутку банку з урахуванням ризику, шляхом підтримання кредитного ризику в межах допустимих значень параметрів. Банки повинні управляти кредитним ризиком, властивим всьому портфелю, а також ризиком окремих кредитів або операцій.

В роботі Базельського комітету існувала традиція відрізняти ринковий ризик від кредитного і враховувати обидві категорії незалежно при розрахунку ризикового капіталу. На практиці позиції в портфелі залежать одночасно від ринкових і кредитних факторів ризику. В цьому випадку апроксимація функції вартості портфеля, що розділяє зміну вартості на чистий ринковий ризик плюс компонент чистого кредитного ризику, може призвести до недооцінювання ризику. Тому не можна стверджувати, що існуючий підхід до регулювання завжди буде консервативним з точки зору оцінювання ризиків. Цей факт часто

обговорюється в контексті позик в іноземній валюті і стверджується, що при традиційному підході до регулювання реальний ризик портфеля позик в іноземній валюті буде значно недооцінений.

Різниця між ринковим і кредитним ризиками та їх незалежним аналізом має певні традиції у банківському регулюванні, особливо в роботі Базельського комітету [18]. Регуляторні органи традиційно вважали, що кредитний ризик в основному важливий для банківської діяльності, а ринковий ризик – в основному для торгової діяльності. Таким чином, нормативна категоризація імітує традиційну організацію банків в кредитний відділ і відділ ринкових інвестицій.

Коли ми залишаємо осторонь операційний ризик, Розділ 1 Базеля II вимагає наявності окремого регуляторного капіталу для кредитного та ринкового ризику [19]:

$$RC_c + RC_m. \tag{1.1}$$

Нормативний капітал для кредитного ризику, *RCc*, сьогодні розраховується для кожної позики окремо. Моделі кредитного ризику портфеля в даний час не використовуються для розрахунку регулятивного капіталу, але вони також підходять для цієї схеми, якщо вони приймають фактори ринкового ризику як детерміновані. Регуляторний капітал для ринкового ризику *RCm* призначений для запобігання виникненню несприятливих змін ринкових цін і не враховує можливість дефолту контрагента. Але для деяких позицій в торговій книзі потрібно мати регуляторний капітал для ризику контрагента Базельським комітетом з банківського нагляду [20].

Проте зв'язок кредитного ризику з банківською книгою і ринкового ризику з торговою книгою, можливо, послужив надихаючим аргументом на користь того, що поточне регулювання, яке виражено формулою (1.1), є консервативним та потребує аналізу таких передумов.

Передумова 1. «Диверсифікація»: при вимірюванні суб-адитивного ризику всього портфеля буде меншою або максимально дорівнювати сумі ризику по банківській і торговельній книгах.

Передумова 2. Кредитний ризик має відношення тільки до банківської книги, а ринковий ризик – до торгової книги.

За всіма показниками суб-адитивного ризику загальний ризик буде меншим або максимально дорівнювати сумі ринкового ризику і кредитного ризику. Це вірний аргумент, якщо передумови виконуються, і висновок обов'язково повинен бути вірним.

Передумова 1 зазвичай виконується за означенням суб-адитивності. Передумова 2 зазвичай не сприймається буквально, але вважається хорошим наближенням до реальності. Регулювання широко вважається консервативним, оскільки вимагає окремого ризикового капіталу для ринкового і кредитного ризику.

Література по інтеграції ринкового та кредитного ризику розглядає проблему інтеграції ризиків по-різному. Існує один напрям, в якому критично розглядається традиційна категоризація. Робота [21] – це рання стаття, в якій розробляється модель за скороченою формою для включення стохастичних процентних ставок в традиційні моделі кредитного ризику. Автори [22] розробляють систему кредитного ризику, яка включає стохастичні процентні ставки, але заснована на структурній моделі кредитного ризику. В роботі [21] пропонується система моделювання для інтегрованого аналізу ринкових і кредитних ризиків для портфелів з фіксованою прибутковістю.

Інша гілка досліджень [22] розглядає інтегроване моделювання ризиків, але має дещо іншу точку зору. Цей підхід не заперечує традиційну класифікацію, а скоріше вказує на те, що портфелі, які аналізуються в рамках різних категорій ринку і кредитного ризику, можна розглядати як ризики підпортфелей для всього банківського портфеля. Ясно, що коли будуть побудовані підпортфелі, то єдиною проблемою, яку ще належить вирішити, є кількісна оцінка ефекту диверсифікації, за умови, що ці підпортфелі будуть об'єднані в загальний портфель. Це саме те, що автори роблять в своїх роботах.

З іншого боку стверджується, що проблема комплексного аналізу ринкових і кредитних ризиків не є проблемою диверсифікації. Проблема часто полягає у тому, що побудова підпортфеля за ринковими та кредитними

факторами ризику неможлива. Якщо це так, то цей факт повинен бути проаналізований в першу чергу. Якщо замість цього в такій ситуації вартість портфеля апроксимується суб-позиторіями ринкового і кредитного ризику, то помилка в оцінюванні складових зазвичай призводить до помилки оцінювання ризику у цілому та до значного заниження справжнього ризику.

1.2 Огляд моделей кредитного ризику

Оцінювання кредитних ризиків – важлива задача загальної проблеми менеджменту ризиків фінансових організацій, які забезпечують клієнтів кредитами. Коректне розв’язання цієї задачі забезпечує повне та своєчасне повернення кредитів і зменшення можливих втрат. Кредитний ризик представляє собою наявний або потенційний ризик для надходжень і капіталу, який виникає через неспроможність сторони, що взяла на себе зобов'язання виконати умови фінансової угоди із банком або в інший спосіб виконати взяті на себе зобов'язання [22].

Під час оцінювання кредитного ризику розрізняють індивідуальний та портфельний ризики. Джерелом індивідуального ризику є окремий, конкретний контрагент банку – позичальник, боржник, емітент цінних паперів. Оцінювання індивідуального ризику передбачає оцінювання кредитоспроможності окремого контрагента, тобто його індивідуальну спроможність своєчасно та в повному обсязі розрахуватися за взятими зобов'язаннями. Такі ризики пов'язані з поверненням кредиту із запізненням або взагалі з його неповерненням.

Статистичні методи оцінювання ризиків найчастіше застосовують для аналізу масових явищ. Ці методи ґрунтуються на інформації, взятої за великий проміжок часу про події, які можуть негативно вплинути на результат. Якщо подібні події мали місце раніше, то визначають, з якою періодичністю вони відбувалися. Також статистичні методи оцінювання ризиків базуються на розрахунках ймовірності того, що настануть ті чи інші події, які можуть мати негативні або позитивні наслідки.

Якісні методи оцінювання ризиків застосовують у тих випадках, які зустрічаються набагато рідше і не можуть бути визначені статистично. Здійснюються вони на підставі знань і досвіду експертів. Тому ці методи оцінювання ризиків ще називають методом експертних оцінок. Рівень ризику залежить від значення (ймовірність настання небажаних подій), яке від нього очікують, і від варіантів можливого результату. У великому потоці кредитних заявок банки намагаються відкинути найменш привабливих і ризикованих позичальників для того щоб приділити час найбільш цікавим за рахунок автоматизації системи управління ризиками, наприклад, за допомогою скорингової моделі і системи, яка нараховує бали в анкеті позичальника і розраховує кредитоспроможність клієнта.

Скоринг – це система оцінювання позичальників, яка заснована математично-статистичних методах. Мета скорингу – оцінити рівень платоспроможності клієнта за певними факторами та відібрати потенційних позичальників. Проблема скорингу актуальна в банківській системі, адже дана модель є основним індикатором оцінювання ризику при прийнятті кредитного рішення. Однак, такі рішення не завжди ідеальні, оскільки система приймає рішення, які можуть не відповідати дійсному стану справ, через що банк може видати кредит неплатоспроможному клієнту або навпаки не видати хорошому позичальнику [23].

Розвиток скорингу та його різновидів почався із створення інформаційних технологій, за допомогою яких можна було б обробляти велику кількість кредитних заявок. Зазвичай підхід, який ґрунтується на скорингу, необхідний для побудови математичних моделей для оцінювання кредитоспроможності позичальників на основі кредитних історій банку та при оцінюванні рівня ймовірності настання дефолту потенційного позичальника, виходячи з його соціально-демографічних показників. Маючи існуючі та проаналізовані статистичні дані «позитивних» і «негативних» кредитів за обраний або встановлений період, банк може визначити безпосередньо ті фактори, які напряму створюють передумови для повернення/неповернення кредиту, та для кожного

нового клієнта на основі цих характеристик визначити його спроможність до повернення кредиту [24]. Початкові етапи скорингового підходу ґрунтуються на експертному висновку, так як першочергово необхідно зібрати та визначити характеристики, які відносяться безпосередньо до клієнта, а також перевірити дані, що були надані самим клієнтом.

Кредитні менеджери під час співбесіди з позичальником збирають максимум інформації про нього, після чого скорингова система обробляє інформацію, нараховуючи бали за кожен сприятливий фактор.

Очевидно, що скорингову модель необхідно корегувати у процесі її використання (кожні 2-3 роки), оскільки ситуація в країнах динамічно змінюється, з'являються певні зловмисники, які можуть прорахувати, які саме фіктивні дані потрібно надати банку для того щоб скорингова модель визначила, що йому можна надати кредит і т. ін. Спільне використання обох підходів надає можливість полегшити процес прийняття рішення стосовно видачі кредиту клієнтам банків [25].

Більш коректним та правильним буде підхід, який базується на основі статистики дефолтів та включає всі попередні періоди, а саме скорингову модель. [25, 26]. Для реалізації цього підходу необхідно передбачити та визначити ймовірність реалізації певних подій з такими вимогами:

– еквівалентними є ті об'єкти для аналізу яких використовуються статистичні дані та для яких ці дані збираються;

– еквівалентними є ті умови, для яких пропонується використовувати та збирати статистичні дані;

– обсяги вибірок статистичних даних є достатніми, методи їх обробки – коректними, а джерело інформації надійним.

Такий підхід висуває високі вимоги до статистики дефолтів, які подані нижче.

– однорідна вибірка (позичальники повинні бути досить схожими);

– вибірка повинна включати певну кількість випадків, притримуючись умови: чим більше в вибірки буде дефолтів - тим краще для результату; як правило, за експертними рекомендаціями та оцінками, обсяг вибірки має становити приблизно 2000 випадків, або більше.

− побудова моделі повинна бути виконана на вибірці, що була накопичена за обмежений термін. Саме зміна макроекономічного середовища призводить до цієї вимоги. Так для одного макроекономічного середовища позичальник з певним набором параметрів буде виплачувати кредит вчасно та у повному обсязі, але в іншому середовищі це може стати причиною класифікувати його як дефолт. Через це вважається, що для країн, які постійно розвиваються, необхідно кожні 2-3 роки змінювати модель оцінювання ймовірності дефолту;

− для отримання коректних результатів щодо кредитування фізичних осіб, необхідно зберігати та просити надати не тільки особисту кредитну історію позичальників (оплачені/неоплачені кредити), а також зберігати інформацію, що стосується віку, статі, місця роботи, посади, сімейного стану тощо. Цю процедуру необхідно виконувати для створення бази знань про клієнта та для розуміння тог, які саме характеристики будуть мати вагу при прийнятті рішення та які з них будуть значущими для побудованої моделі. Через це на перших етапах необхідно накопичити якомога більше характеристик та параметрів по кожному клієнту;

− вибірка має включати інформацію по кредитах, цикл кредитування яких вже закінчився. Ця вимога є необхідною, оскільки обов'язковою є інформація про те, чи був цей кредит повернений, чи ні [27];

− до кредитної історії зазвичай відносять такі види продуктів (споживчий кредит, кредит на авто, іпотечний кредит). Для кожного такого кредитного продукту повинна бути побудована скорингова модель [28].

Дані стосовно процесу кредитування фізичних осіб задовольняють практично всім цим вимогам. Для фізичних осіб параметрами скорингової моделі можуть бути вік, сімейний стан, кількість дітей, освіта, місце проживання, робота, посада, власність, кредитна історія, наявність і розмір поточних боргів, тривалість відносин з кредиторами, співвідношення кількості поданих заявок і виданих кредитів та інше [28].

1.3 Регресійний аналіз та регресійні моделі

Регресійний аналіз — розділ математичної статистики, присвячений методам аналізу залежності однієї величини від іншої. Такий підхід використовується в тому випадку, якщо відношення між змінними можуть бути виражені кількісно у виді деякої комбінації цих змінних. Отримана комбінація використовується для передбачення значення, що може приймати цільова (залежна) змінна, яка обчислюється на заданому наборі значень вхідних (незалежних) змінних. У найпростішому випадку для цього використовуються стандартні статистичні методи, такі як лінійна регресія [29].

1.3.1 Регресійні моделі

Регресійні моделі ґрунтуються на виявленні причинно-наслідкових зв'язків між спостережуваними індикаторами і рівнем ризику. Розрізняють дві основні групи показників, які можна використовувати як спостережувані індикатори (пояснювальні змінні): 1) змінні оточення – кількісні показники; 2) фактори ризику, тобто кількісні показники, що характеризують спостережувані випадки реалізації ризиків [29]. Така математична модель має вигляд:

$$x = A \times f + b + \varepsilon \tag{1.2}$$

де x – величина втрат, пов'язаних з кредитним ризиком; f – вектор значень спостережуваних змінних; ε – випадкова величина, що визначає рівень похибки моделі; A і b – оцінювані параметри, що характеризують залежність між зміною x та факторними змінними f. Для застосування цього методу необхідно мати достатній обсяг даних з метою забезпечення високої точності оцінок [30].

1.3.2 Лінійна ймовірнісна модель

Лінійна імовірнісна модель – це в сутності регресійна модель, де залежна змінна набуває значення 0 або 1, де 0 означає, що кредит затверджено, 1 – не

затверджено. З математичної точки зору лінійна модель виражається таким правилом прийняття рішення:

$$y = \beta_0 + \beta_1 x_1 + ... + \beta_k x_k + \varepsilon, \tag{1.3}$$

де y – залежна пояснювана змінна; $(x_1, x_2, ..., x_k)$ – незалежні (екзогенні) змінні; $(b_1, b_2, ..., b_k)$ – вагові коефіцієнти, присвоєні незалежним змінним; ε – випадкова похибка, розподіл якої залежить від значень екзогенних змінних, при цьому її математичне очікування має дорівнювати нулю.

У векторному представленні модель має вигляд:

$$y = b' \times x + \varepsilon, \tag{1.4}$$

де x – вектор екзогенних змінних; b' – транспонований вектор параметрів.

Після оцінювання коефіцієнтів моделі її можна використовувати для оцінювання кредитоспроможності для майбутніх запитів [30]. Під час прийняття остаточного рішення стосовно підтвердження запиту на позику, варто брати до уваги межу (поріг) відсічення. Варто зазначити, що позитивною стороною лінійної моделі є простота її реалізації і подальших розрахунків, в той час як якість моделі не завжди є прийнятною. Отже, у процесі побудови моделі лінійної регресії знаходять оцінки вектора параметрів $(b_1, b_2, ..., b_k)$ з використанням експериментальних значень y і $(x_1, x_2, ..., x_k)$; а надалі цю модель застосовують для оцінювання кредитоспроможності клієнта.

1.3.3 Логіт і пробіт моделі

Недосконаліть лінійних методів прогнозування спонукала до пошуку та аналізу більш ефективних методів. Очевидно, що недосконаліть полягає в тому, що оцінена ймовірність не обов'язково знаходиться в інтервалі [0; 1]. Проблему можна вирішити, якщо вдасться знайти відповідне перетворення, що гарантує знаходження оцінки в цьому інтервалі [30]. У вище описаній моделі залежна

змінна *y* є бінарною, тобто приймає тільки два значення і являється функцією параметрів особи-заявника.

Кумулятивні функції розподілу являють собою набір перетворень, що переводять значення ймовірності в інтервал [0; 1]. Також ці функції мають властивість монотонності, - або монотонно зростають, або монотонно спадають. Припустимо, що для вираження ймовірності можна використати стандартний нормальний розподіл:

$$p = \Phi(b'x) = \int_{-\infty}^{b'x} \varphi(z)dz, \tag{1.5}$$

де $\varphi(z)$ – функція щільності нормального розподілу.

У випадку, коли ймовірність описується логістичною функцією розподілу, то отримаємо логіт-модель.

$$p = \Phi(b'x) = \int_{-\infty}^{b'x} \varphi(z)dz = \frac{1}{1+e^{-b'x}}, \tag{1.6}$$

або

$$p = \frac{e^{b_1x_1+\ldots+b_kx_k}}{1+e^{b_1x_1+\ldots+b_kx_k}} \tag{1.7}$$

В більшості випадків оцінка вагових коефіцієнтів для обох моделей відбувається з використанням методу максимальної правдоподібності, реалізація якого є відносно простою і не вимагає великих обчислювальних витрат, а також дозволяє проводити оцінювання параметрів моделі в автоматичному режимі.

1.3.4 Метод на основі нечіткої логіки

Підхід на основі нечіткої логіки дозволяє найкращим чином застосувати експертну оцінку для проведення аналізу ризиків у випадках, коли початкові дані наявні у неповній формі, або взагалі відсутні. Нечітка логіка наближує модель до міркувань людини в процесі прийняття і обґрунтування рішень. Методи нечіткої логіки можна застосовувати для оцінювання обсягу втрат і виявлення ризику банкротства. На зарубіжних ринках використовують автоматизовані системи з

використанням методів нечіткої логіки, що дозволяють оцінювати фінансові ризики.

1.3.5 Метод на основі экспертного оцінювання

Бувають випадки, коли при дослідженні дуже складних систем, у тому числі і фінансових, можуть виникати проблеми, які знаходяться поза межами формальних математичних задач. Ідеєю цього методу є використання інтелекту людей та їх здатностей у знаходженні рішень для слабко формалізованих задач.

Методика проведення експертного оцінювання включає в себе такі етапи:

- формулювання цілей оцінювання;
- постановка задачі;
- створення групи управління процесом оцінювання;
- опис форми отримання необхідних результатів;
- вибір експертів та визначення їх компетентності.

Вибір експертів здійснюється таким чином, щоб ті фахівці, які входять до групи, були ознайомлені та розумілися із специфіці задач (об’єктом прийняття рішень); а також не були б зацікавленні в результатах оцінювання. Тому до групи експертів включають, як правило, 2–3-х фахівців із своєї організації і 2–3-х зовнішніх експертів. Професійна та наукова діяльність експертів дозволяє об’єктивно визначити ступінь їх компетентності та кваліфікації у певній області знань. Суб’єктивний метод оцінювання компетентності полягає у взаємному оцінюванні кожним експертом кваліфікації своїх колег і своєї за певною шкалою. Після оброблення результатів опитування встановлюється компетентність експертної групи, яка визначає можливу похибку оцінювання [29].

Експертний підхід передбачає, що фахівці в області кредитування індивідуальних позичальників визначають суттєві характеристики клієнта банку, які можуть впливати на повернення кредиту, ставлять у відповідність цим характеристикам певні ваги. Для кожного клієнта за всіма характеристиками проставляються бали відповідно до вагових коефіцієнтів, встановленими експертами, та підраховується сума всіх балів. Для банку завчасно

встановлюється певне порогове значення, яке визначає правило: якщо сума всіх балів є меншою цього значення, то клієнту не слід видавати кредит, а якщо більшою цього значення, то клієнт може отримати кредит. Також банк може встановити ще одне обмеження (верхня межа), яке визначає таке правило: якщо сума балів є більшою встановленого значення, то клієнту можна видавати кредит без надмірної обережності. Якщо і верхня і нижня межа щодо видачі кредиту задані, то значення між цими межами знаходяться у компетенції менеджера банку, який на власний розсуд визначає – видавати кредит клієнту, чи ні [30, 31].

1.3.6. Байєсівський підхід на основі мереж довіри

Байєсівські мережі довіри (БМ) дають можливість відобразити причинно-наслідкові зв'язки між різними впливовими чинниками ризику і змінами середовища. На відміну від регресійних моделей, байєсівські мережі дають можливість враховувати не лише безпосередні залежності рівня ризику від факторів ризику, а й залежності між факторами ризику. Крім того, цей клас моделей надає більше можливостей для формування висновку на основі неповних даних [32]. З математичної точки зору БМ – орієнтований граф, вершинам якого відповідають чинники ризику і зміни середовища, а ребрам — виявлені або передбачувані взаємозв'язки. Мережа також описується множиною умовних розподілів випадкових величин, що характеризують чинники ризику і змінні середовища. При використанні байєсівських мереж як інструменту аналізу даних вирішують дві математичні задачі: (1) побудови структури БМ, і (2) формування ймовірнісного висновку. Задача побудови БМ за заданими навчальними даними є NP-складною, тобто задачею нелінійної поліноміальної складності.

1.4 Моделі ринкового ризику

Ринковий ризик – це ризик втрати позицій, що виникають внаслідок руху ринкових цін. Не існує єдиної класифікації, оскільки кожна класифікація може

стосуватися різних аспектів ринкового ризику. Але найбільш часто використовуваними ринковими ризиками є:

- ризик власного капіталу, ризик зміни цін на акції або акції;
- процентний ризик, ризик зміни процентних ставок або їх нестабільності (за припущенням);
- валютний ризик або ризик зміни валютних курсів;
- товарний ризик – ризик, пов'язаний із зміною цін на товари;
- маржинальний ризик є наслідком невизначених майбутніх відтоків грошових коштів, викликаних маржинальними викликами, що покривають несприятливі зміни цінності даної позиції.

Ринковий ризик відносять до ризику втрат в торговельній книзі банку через зміни цін на акції, процентні ставки, кредитні спреди, курси валют, цін на сировинні товари та інших показників, значення яких встановлюються на публічному ринку. Для аналізу та менеджменту ринкового ризику на сучасному етапі банки застосовують ряд досить складних математичних і статистичних методів, які забезпечують створення адекватних моделей та обчислення високоякісних оцінок.

1.4.1 Value-at-Risk

Value-at-Risk (VaR) – це кількісна міра потенційного збитку (у грошовому вираженні) за позиціями фактичної вартості активів внаслідок руху ринку, яка не буде перевищена протягом певного періоду часу і з певним рівнем достовірності.

Важливість цього ризику для торгових підприємств обгрунтовується їх власною внутрішньою моделлю. Базуючись на рекомендаціях [33], було затверджено внутрішню модель для розрахунку регулятивного капіталу для загальних і специфічних ринкових ризиків. Запропонована модель постійно удосконалюється.

VaR характеризується трьома, поданими нижче, параметрами.

- Часовий горизонт, який залежить від ситуації, що розглядається. За Базельським документом це 10 днів, за іншими методиками – 1 день. Найчастіше

поширений розрахунок з часовим горизонтом 1 день. 10 днів використовують для розрахунку величини капіталу, що покриває збитки.

– Довірчий рівень (confidence level) – рівень допустимого ризику. За Базельським документом використовується величина 99%.

– Базова валюта, в якій вимірюється показник ризику.

За останні 15 років методика аналізу VaR стала галузевим і нормативним стандартом для вимірювання ринкового ризику. Вимоги, що пред'являються до VAR та інших методів надзвичайно зросли завдяки новим продуктам, таким як кореляційна торгівля, опціони з декількома активами, реверсивні двовалютні свопи, номінальна вартість яких непередбачувано амортизується, і десятки інших подібних інновацій.

Фактори ризику, необхідні для аналізу ціноутворення в торговельній книзі глобальних установ, виросли до декількох тисяч, а іноді і до 10 000. Моделі оцінювання можливих втрат стають все більш складними і більшість банків зараз перебувають в процесі інтеграції нової аналітики стрес-тестування, яка може передбачати широкий спектр макроекономічних змін.

Незважаючи на існуючі досягнення, VAR та інші методики анадізу ризику часто призводять о некоректних результатів. Криза 2008 року продемонструвала обмеження моделювання ризиків [34]. У 2011 році, незадовго до початку європейської суверенної кризи, моделі ризиків багатьох банків розглядали державні облігації Єврозони як практично безризикові.

Було відзначено залежність VAR від нормального розподілу ринкових процесів і фундаментальне припущення про те, що позиції можуть бути легко ліквідовані. Регуляторні органи намагалися компенсувати деякі з цих обмежень, зокрема, через Базель II.5, зроблено комплексне оновлення системи ринкових ризиків, яка вступила в силу в грудні 2011 року. Деякі нові елементи в структурі методики, такі як вимога для розрахунку стресового значення VAR, підвищують зважені з урахуванням ризику активи (RWA) і підвищують вимоги до об’єму резервного капіталу в два-три рази.

Введення більш високих вимог до капіталу може зробити фінансову систему більш безпечною, але з точки зору моделювання це досить недосконалий інструмент. Тривалі удосконалення в стрес-тестуванні є корисним доповненням до основної методики VAR, але майже всі банки згодні з тим, що моделі оцінювання ризику вимагають подальшого удосконалення [35]. Банки цікавляться варіантами, пов'язаними з адекватним моделюванням і оцінюванням; вони шукають коректний баланс між системністю і точністю оцінювання з одного боку, а також простотою, прозорістю і швидкодією процедур з іншого. Наявність високоякісних ринкових даних виявляється настільки ж важливою, як і самі моделі, але багато банків не впевнені, де провести межу між прийнятним і неприйнятним рівнями якості. Існує кілька способів оцінювання VaR:

− параметричний: оцінка виконується в припущенні, що відомий тип розподілу прибутків (найчастіше він передбачається нормальним або логнормальним);

− історичний (непараметричний): розподіл прибутку береться з уже реалізованого часового ряду, тобто передбачається, що розподіл прибутку в майбутньому буде аналогічний історичному;

− метод Монте-Карло.

На практиці зазвичай використовують два параметричних методи розрахунку VaR: дельта-нормальний VaR та дельта-гамма VaR.

Найбільш популярним параметричним методом розрахунку Value-at-Risk, є дельта-нормальний метод [35, 36]. При розрахунку Value-at-Risk за дельта-нормальним методом використовуються припущення про нормальність розподілу всіх ринкових чинників, що впливають на вартість портфеля і про лінійний зв’язок між змінами факторів ризику і фінансовими результатами за складовими портфеля. В цьому випадку, результат по портфелю буде являти собою суму нормально розподілених величин, тобто теж нормально розподілену величину [36].

Значення Value-at-Risk згідно дельта-нормальному методу може бути розраховане згідно (1.8):

$$VaR = K\sqrt{\sum_{i,j=1}^{N} \sigma_{ij} D_i D_j}\,, \qquad (1.8)$$

де D_i – чутливість (дельта) портфеля до *i*-го фактору ризику (сума коефіцієнтів лінійного зв'язку за *i*-м фактором результатів за всіма складовими портфеля); K – коефіцієнт, що залежить від обраного довірчого інтервалу (показує у скільки разів втрати для заданої довірчої ймовірності більше стандартного відхилення нормального розподілу); σ_{ij} – коваріація *i*-го і *j*-го факторів ризику; N – кількість факторів ризику.

Таким чином, для використання даного методу необхідно знати матрицю коваріацій ринкових факторів (волатильності окремих факторів будуть враховані в даній матриці як коваріація фактора самого з собою). Ця матриця може бути отримана як на основі історичних даних, так і на основі прогнозів. Двома найбільш часто використовуваними значеннями коефіцієнта K є такі: 2,33 (для ймовірності 0,99) і 1,65 (для ймовірності 0,95).

Дельта-нормальний метод має такі переваги: відносна простота реалізації; швидкість обчислень; дозволяє використовувати різні варіанти значень волатильності та кореляцій.

Недоліки дельта-нормального методу: неможливість використання інших розподілів, крім нормального, в силу чого не враховуються "важкі хвости"; неможливість коректного врахування ризиків при використанні нелінійних інструментів; імовірність появи суттєво значущих похибок у використовуваних моделях [36].

Найскладнішим методом, який використовують для розрахунку VaR є метод Монте-Карло. Незважаючи на складність, даний підхід може дати точність вище, ніж за іншими методами. З використанням методу Монте-Карло можна передбачити велику кількість випробувань - побудову разових моделей для розвитку ситуацій на ринку із розрахунком результатів по портфелю. Результатом проведення даних випробувань є отримання розподілу можливих

фінансових результатів, в тому числі і втрат, на основі якого може бути отримана VaR-оцінка шляхом відсікання найгірших результатів для обраної довірчої ймовірності.

Використання методу Монте-Карло, як і для параметричного VaR, передбачає побудову таких моделей:

- модель залежності вартості фінансового результату по портфелю від змін факторів ризику;
- модель волатильності і кореляцій факторів ризику.

Метод Монте-Карло не передбачає узагальнення формул для отримання аналітичної оцінки портфеля в цілому, а тому для отримання результату по портфелю, а також для волатильності і кореляцій можна використовувати значно складніші моделі. Оскільки оцінювання VaR за методом Монте-Карло практично завжди проводиться з використанням програмних засобів, то такі моделі реалізуються досить складними підпрограмами. Тобто метод Монте-Карло дозволяє використовувати при розрахунку ризиків моделі практично будь-якої складності [36].

Переваги методу Монте-Карло такі: можливість розрахунку ризиків для нелінійних інструментів; можливість використання будь-яких розподілів ймовірностей; можливість моделювання складної поведінки ринків – трендів, кластерів високої або низької волатильності, нестаціонарних кореляцій між факторами ризику, сценаріїв типу "що-якщо" і т.д.; можливість подальшого, практично нічим не обмеженого, розвитку моделей [36].

Недоліки методу Монте-Карло: висока складність реалізації; вимогу до потужності обчислювальних ресурсів; імовірність наявності значущих помилок у використовуваних моделях.

1.4.2 Stressed Value-at-Risk

Стресова VaR розраховує вартість під ризиком на основі наявності значного стресу на ринку протягом одного року. Розраховується величина такого ризику при використанні рівня достовірності 99%. Термін зберігання становить один

день для внутрішніх цілей і десять днів для цілей регулювання. Історичні ринкові дані і спостереження за кореляцією за період значного фінансового стресу (тобто процеси характеризуються високою волатильністю) використовуються в якості вихідних даних для моделювання по методу Монте-Карло.

Процес вибору часового вікна для розрахунку стресового значення ризику грунтується на ідентифікації вікна, що характеризується високим рівнем волатильності серед основних факторів оцінювання ризику. Стрес-тестування в управлінні ризиками часто включає в себе дуже складні, комп'ютерно-генеровані імітаційні моделі, які використовують різні сценарії в якості тестової бази; при цьому аналізується, яким чином баланс організації відповідає конкретним ситуаціям.

Наприклад, у період невизначеності фінансові організації можуть розгортати ці моделі для аналізу ринкового та портфельного ризиків і прийняття обґрунтованих рішень на основі результатів. Наявність високоякісних даних у своєму розпорядженні після виконання стрес-тестів в управлінні ризиками дає можливість фінансовим організаціям бути більш ефективними, пом'якшувати ризики та виявляти проблеми на ранніх стадіях.

Надмірне сприймання ризиків великими банками, відсутність прозорості, неспроможність фінансового регулювання та суб-стандартне кредитування призвели до катастрофічної ситуації, яка в свою чергу призвела до значних змін у фінансовому регулюванні. Прямим результатом цього було те, що промисловість стала свідком значних зрушень у банківському регулюванні та нагляді, особливо щодо регуляторної звітності та звітності стосовно менеджменту ринкових ризиків. Система правил Базель II була переглянута і посилена рамками Базеля III. Базель III був розроблений з метою розширення регуляторної сфери управління ризиками; він гарантує фінансовим організаціям юридичне право проводити комплексне стрес-тестування. Це дозволяє їм управляти ризиками і визначати їх здатність виживати в конкретних фінансових умовах.

З липня 2018 року, згідно з Базельськими правилами, банки повинні проводити стрес-тестування згідно [37]. Крім того, вони також повинні повністю документувати, яким чином були отримані результати стрес-тестів. Згодом для більш ефективного управління процесом стрес-тестування, пом'якшення ризиків і забезпечення постійного дотримання нормативних вимог фінансовим організаціям потрібне єдине, скоординоване рішення стрес-тестування для управління балансом.

Переваги рішення для стрес-тестування. Оскільки нові вимоги щодо стрес-тестування, пов'язані з управлінням, часто оновлюються, існує чітка потреба в комплексному стрес-тестуванні для вирішення проблем управління ризиками. Для великих фінансових організацій стрес-тестування навряд чи є новим. Новими є рівні складності та прозорості, які вимагають ці тести.

Чинні положення та вимоги до надійного стрес-тестування показують, що фінансові організації повинні бути послідовними та прозорими з отриманням необхідних результатів. Це означає, що централізація даних, надання даних, їх консолідація та агрегація стали ключовими аспектами, які необхідно враховувати для задоволення нормативних вимог.

Проблеми, з якими стикаються деякі фінансові організації у прагненні прозорості та точності, – це старі системи, неефективна інфраструктура та проблеми, пов'язані з впровадженням та звітуванням про моделі ризику. Старі системи можуть призводити до використання непослідовного визначення даних, що ускладнює їх консолідацію та розробку послідовного і повторюваного процесу їх аналізу. Неефективна інфраструктура призводить до відключення деяких процесів протягом всього життєвого циклу моделі, що призводить до невідповідності результатів стрес-тестування. Нарешті, фінансові організації можуть не мати добре структурованого, документованого та прозорого процесу, який відповідає нормативним вимогам. Це означає, що результати стрес-тестування часто є помилковими та фрагментованими.

Проте, маючи рішення стосовно стрес-тестування, управління ризиками може бути проаналізовано на одній централізованій платформі. Фінансові

організації можуть виконувати конкретні та індивідуальні фінансові сценарії, які узгоджуються з останніми правилами менеджменту ризиків. Такі сценарії використовують для аналізу тверджень «що-якщо», оцінювання ризику портфеля та планування управління капіталом.

Використовуючи рішення, отримані в результаті стрес-тестування, всі учасники можуть працювати з однаковим, повністю інтегрованим і узгодженим джерелом даних, яке включає створення логічних структур даних та їх належну ієрархію, забезпечуючи надійність, точність і повторюваність стрес-тестування. І в будь-який момент процесу стрес-тестування інформація є легко доступною для пошуку та аналізу, забезпечуючи необхідну прозорість.

1.5 Програмні засоби СППР. Огляд інструментальних засобів

Даний підрозділ присвячено програмним засобам та мовам програмування для побудови моделей оцінювання кредитоспроможності клієнтів. Основними існуючими конкуруючими програмними рішеннями в області побудови моделей на сучасному ринку інформаційних технологій є такі програмні засоби:

1. SAS Institute або SAS Institute Inc. Моделювання та оцінювання кредитів є важливою складовою оцінки вимог до капіталу, і банки стикаються з різними проблемами та потребами, пов'язаними з цим моделюванням. SAS® Credit Scoring for Banking – це комплексне рішення, яке дозволяє детально проаналізувати та покращити прогнозування кредитного ризику, враховуючи ці виклики та потреби [38].

Програма заснована на досвіді, отриманому від впровадження SAS® Credit Scoring для банківської діяльності для ряду банків. Огляд архітектури підкреслює програмне забезпечення SAS®, що входить до SAS® Credit Scoring for Banking, та етапи, на яких вступає в дію різне програмне забезпечення. Підхід SAS® Credit Scoring для банківської діяльності оцінює її масштабованість стосовно додаткових джерел даних та моделей. SAS® Credit Scoring for Banking забезпечує результат, що містить усі необхідні компоненти для оцінювання кредитів: вилучення та

агрегування даних, створення змінних, розробка моделі та форми для модельної звітності (рис. 1.1).

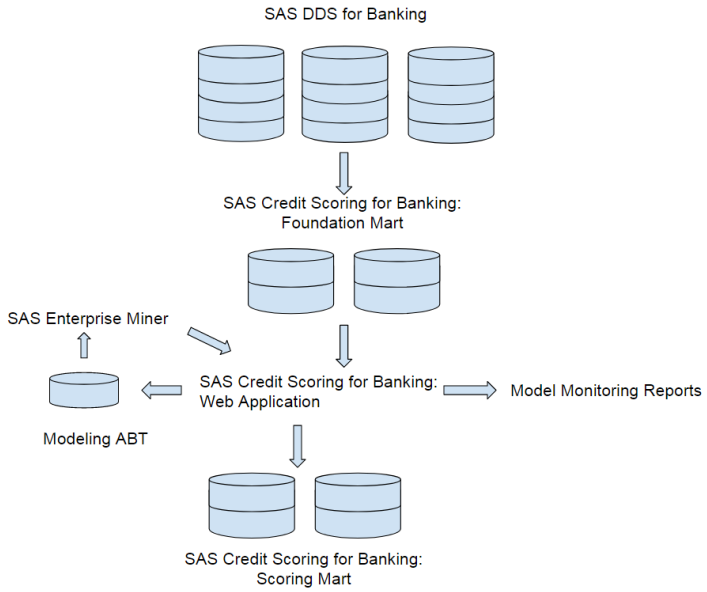

Рисунок 1.1 – Загальна схема вилучення та агрегації даних

На схемі вище представлено огляд архітектури високого рівня SAS® Credit Scoring для банківської діяльності. Extract Load Transform (ETL) завантажують відповідні дані для підрахунку кредитів із сховища даних та завантажують дані у Foundation Mart.

2. FICO™ Model Builder's Scorecard module. Модуль показників FICO Scorecard допомагає отримати уявлення про свої дані та прогнозні відносини всередині них, а також вирішити проблеми моделювання, які, найімовірніше, виникнуть у практиці оцінювання балів. За допомогою модуля Scorecard можна створювати високо передбачувані показники без шкоди для операційних чи правових обмежень і легко розміщувати ці моделі в операціях, починаючи з короткого вступу до підрахунку балів та аналізу його відношення до статистичного моделювання (рис. 1.2). До них відносяться формули балів та інженерія підрахунків, вирівнювання, підгонка цілей та алгоритмів оцінювання, калібрування та масштабування балів, формування висновку про ефективність, перевірка завантажувальної версії [39].

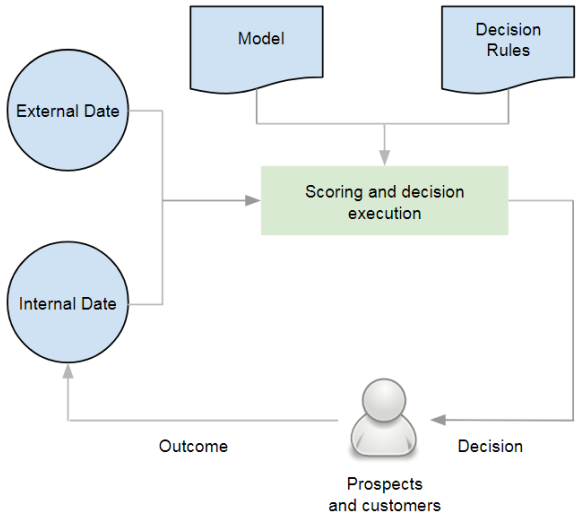

Рисунок 1.2 – Загальна архітектура програми Builder’s Scorecard для прогнозування

Вище представлена архітектура FICO™ Model Builder’s Scorecard та її модулів, які відповідають за аналіз, обробку даних та прогнозування.

3. IBM SPSS Modeler. Рішення IBM, що може використовуватись для побудови ефективних скорингових моделей. IBM SPSS Modeler – це програмне забезпечення для обміну даними та текстової аналітики від IBM. Використовується для побудови прогнозних моделей та виконання інших аналітичних завдань. Має візуальний інтерфейс, який дозволяє користувачам використовувати алгоритми статистики та обробки даних без програмування (рис. 1.3). Однією з його головних цілей з самого початку було позбутися зайвої складності в трансформації даних та зробити складні прогнозні моделі дуже зручними у використанні. Перша версія включала дерева рішень (ID3) та нейронні мережі (backprop), які можна було б навчити.

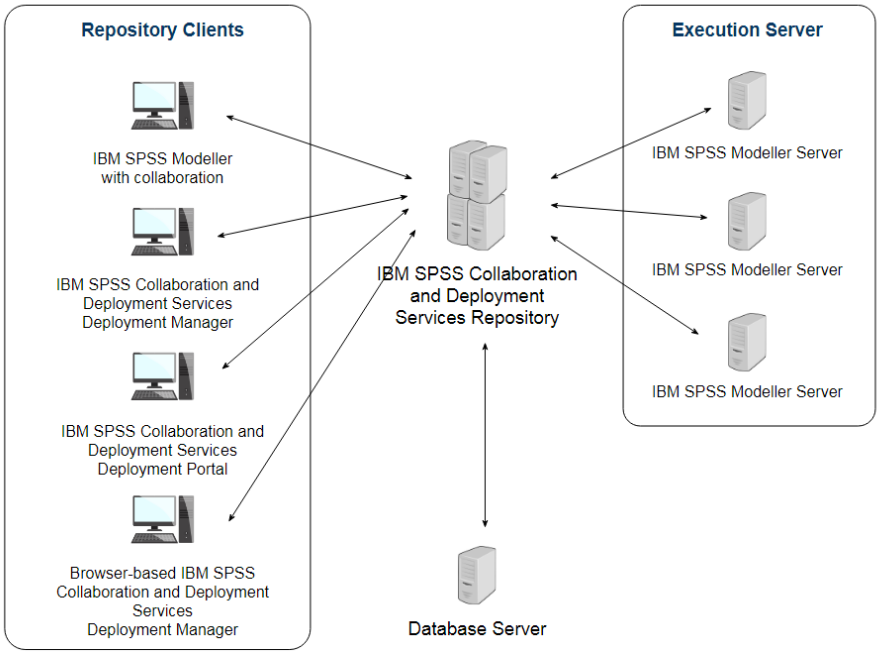

Рисунок 1.3 – Архітектура модулів IBM SPSS Modeler

4. SAP SE – Системи, додатки та продукти в обробці даних. Європейська багатонаціональна корпорація програмного забезпечення, яка виробляє корпоративне програмне забезпечення для управління бізнес-операціями та відносинами з клієнтами. Компанія особливо відома своїм програмним забезпеченням ERP.

Це набір прогнозного моделювання, який допомагає фахівцям-аналітикам та керівникам підприємств отримувати інформацію з даних. Серед інших функцій InfiniteInsight використовується для виявлення важливих змінних, класифікації, регресії, сегментації, часових рядів, рекомендацій щодо продукту, як описано та виражено інтерфейсом Java Data Mining, та для аналізу соціальних мереж. InfiniteInsight дозволяє передбачити поведінку чи значення, прогноз часового ряду або розуміння групи осіб із подібною поведінкою. Додаткові функції включають поведінкове моделювання, експорт модельного коду в різні цільові середовища або побудову прогнозних моделей поверх файлів даних SAS або SPSS. Конкурентами є SAS Enterprise Miner, IBM SPSS Modeler та Statistica. Інструменти прогнозування з відкритим кодом, такі як пакет R або Weka, також є конкурентами, оскільки вони надають подібні функції безкоштовно.

Окрім готових програмних засобів, які можна застосовувати для аналізу кредитоспроможності, існують мови програмування, на яких написана велика кількість бібліотек та фреймворків з відкритим кодом, що дозволяють спрощувати розробку та написання програм з нуля.

1.5.1 Мова програмування Python

Python – високорівнева мова програмування загальнодоступних даних, орієнтований на підвищення продуктивності та швидкість роботи. Синтаксис ядра Python мінімальний. У той час, коли стандартна бібліотека включає великий об'єм корисних функцій. Python підтримує технологічне, об'єктно-орієнтоване, функціональне та логічно орієнтоване програмування. Основні архітектурні характеристики – динамічна типізація, автоматичне управління пам’яттю, повна інтроспекція, механізм обробки підключень, підтримка багатопоточних обчислювань, високорівневі основні структури даних. Підтримується розбирання програмного забезпечення в модулях, які, у свою чергу, можуть бути об'єднані в пакети [40].

Еталонна реалізація Python – це інтерпретатор CPython, що підтримує велику кількість активно використовуваних платформ. Він розповсюджується під ліцензією ліцензії Python Software Foundation, яка використовує його без обмежень у будь-яких додатках [41]. Є реалізація інтерпретатора для JVM з можливістю компіляції CLR, LLVM та іншими незалежними реалізаціями. Сьогодні активно розвиваються мовні програми, нові версії з додаванням / зміною мовних власних результатів, які виходять приблизно один раз на два з половиною роки.

Їх основні бібліотеки такі:

1. NumPy – це бібліотека мови Python, що додає підтримку великих багатовимірних масивів і матриць, разом з великою бібліотекою високорівневих математичних функцій для операцій з цими масивами. Математичні алгоритми, реалізовані на Python, часто працюють набагато повільніше тих же алгоритмів, реалізованих на компільованих мовах (наприклад, Фортран, Сі, Java). Бібліотека

NumPy надає реалізації обчислювальних алгоритмів (у вигляді функцій і операторів), оптимізовані для роботи з багатовимірними масивами [40, 41].

2. SciPy – це відкрита бібліотека високоякісних наукових інструментів для мови програмування Python. SciPy містить модулі для оптимізації, інтегрування, спеціальних функцій, обробки сигналів, обробки зображень, генетичних алгоритмів, розв’язування звичайних диференціальних рівнянь та інших задач, які зазвичай вирішуються в науці і в інженерних розробках. Для візуалізації при використанні SciPy часто застосовують бібліотеку Matplotlib, що є аналогом засобів виведення графіки MATLAB. Сьогодні SciPy поширюється під ліцензією BSD і його розробники спонсоруються Enthought [41].

3. Pandas – це бібліотека Python, яка є потужним інструментом для аналізу даних. Пакет дає можливість будувати зведені таблиці, виконувати групування даних, надає зручний доступ до табличних даних, а за наявності пакета matplotlib дає можливість рисувати графіки на отриманих наборах даних.

Основні можливості бібліотеки: об'єкт DataFrame для маніпулювання індексованими масивами двовимірних даних; інструменти для обміну даними між наявними структурами в пам'яті і файлами різних форматів; засоби інтегрування даних і способи обробки пропусків; переформатування наборів даних, в тому числі створення зведених таблиць; зріз даних за значеннями індексу, розширені можливості індексування, вибірка з великих наборів даних; вставка і видалення стовпців даних; можливості групування дозволяють виконувати триетапну операцію типу «поділ, модифікація, об'єднання»; злиття і об'єднання масивів даних; ієрархічне індексування дозволяє працювати з даними високої розмірності в структурах меншої розмірності; робота з часовими рядами: формування часових періодів і зміна інтервалів.

4. Matplotlib – це бібліотека Python для побудови якісних двовимірних графіків. Matplotlib є гнучким, легко конфігурованим пакетом, який разом з NumPy, SciPy і IPython надає можливості, подібні MATLAB. В даний час пакет працює з кількома графічними бібліотеками, включаючи wxWindows і PyGTK [40]. Пакет підтримує багато видів графіків та діаграм: графіки (line plot);

діаграми розкиду (scatter plot); стовпчасті діаграми (bar chart) і гістограми (histogram); кругові діаграми (pie chart); лист діаграми (stem plot); контурні графіки (contour plot); поля градієнтів (quiver); спектральні діаграми (spectrogram). Користувач може вказати осі координат, додати написи і пояснення, використовувати логарифмічну шкалу або полярні координати. Нескладні тривимірні графіки можна будувати за допомогою набору інструментів (toolkit) mplot3d. Є й інші набори інструментів: для картографії, для роботи з Excel, утиліти для GTK і інші [41].

5. Scikit-learn це бібліотека для машинного навчання на мові програмування Python з відкритим кодом. За допомогою бібліотеки можна реалізувати алгоритми класифікації, регресії і кластеризації, в тому числі алгоритми SVM, випадкового лісу, k-найближчих сусідів і DBSCAN, які побудовані на взаємодії бібліотек NumPyi SciPy з Python. Перевагами даної бібліотеки є такі: прості та ефективні інструменти для data mining і data analysis; зручний доступ до необхідних компонентів; побудований на NumPy, SciPy і Matplotlib; відкритий вихідний код, ліцензія BSD.

6. Keras – це бібліотека, що дозволяє на більш високому рівні працювати з нейромережами. Вона спрощує безліч завдань, використовується в швидких експериментах і суттєво зменшує об’єм одноманітного коду. Як бекендна бібліотека для обчислень keras може використовувати theano і tenzorfow.

1.5.2 Мова програмування R

R – інтерпретована мова програмування, основним способом роботи з яким є командний інтерпретатор. Мова є чутливою до регістру, в плані синтаксису схожа, з одного боку, на функціональні мови типу Scheme, з іншого – на типові сучасні сценарні мови, з простим синтаксисом і невеликим набором основних конструкцій. Мова об'єктна: будь-який програмний об'єкт в ній має набір атрибутів – іменований список значень, що визначають його. Мова R широко використовується серед статистиків та аналітиків даних для розробки статистичного програмного забезпечення і аналізу даних.

Пакет GNU – вихідний код для програмного середовища R, написано на C, Fortran і R; він є у вільному доступі за загальною публічною ліцензією GNU. Попередньо складені двійкові версії передбачені для різних операційних систем. Хоча R має інтерфейс командного рядка, є кілька графічних інтерфейсів користувача, таких як RStudio, інтегроване середовище розробки.

Мова підтримує мінімальний набір примітивних типів даних: символьний (character), числовий (numeric), логічний (logical) і комплексний (complex). Числові змінні, крім звичайних чисел, можуть приймати спеціальні значення NaN (Not a Number – "не число») і Inf (Infinity – «нескінченність»). Нескінченність (додатна чи від’ємна) виходить при виході результату обчислень за межі уявного реалізацією діапазону; NaN – при операціях з невизначеним результатом. Крім цих, є ще одне дуже важливе спеціальне значення, NA (Not Available – «не доступне»).

Значення примітивних типів можуть об'єднуватися у вектори (vector), списки (list), матриці або масиви (matrix), в тому числі багатовимірні; ці комбіновані типи зберігають набори даних одного і того ж примітивного типу. Крім цього мова містить поняття факторів (factor) – наборів категоріальних або шкальних даних, які приймають строго певний набір значень. Можуть створюватися таблиці (data frame) – структури даних, які для кожного рядка (індивіда) зберігають набір різних (і мають різні типи) параметрів (ознак).

Функції R можуть об'єднуватися в пакети – завантажувальні модулі, які підключаються до будь-якої програми і надають об'єднані в них обчислювальні засоби. Пакети для R можуть розроблятися на інших мовах програмування, в тому числі на C, що дозволяє, з одного боку, компенсувати обмеженість образотворчих засобів самої мови R, а з іншого – за необхідності досягти високих показників обчислювальної продуктивності.

Сама мова має досить обмежені і не дуже зручні засоби опису даних, але це компенсується наявністю бібліотечних засобів, які дозволяють завантажувати в вигляді таблиць R вибірки даних, представлені у більшості відкритих форматів.

Так, в R можуть бути легко завантажені таблиці в простому текстовому форматі, таблиці Excel різних версій, дані в форматах CSV, XML і багатьох інших.

CRAN – це мережа ftp та веб-серверів у всьому світі, які зберігають однакові, сучасні версії коду та документації для R. Сьогодні CRAN містить 8341 пакетів. Крім CRAN, є й інші репозиторії з великою кількістю пакетів. Синтаксис для установки будь-якого з них простий: install.packages ("Name_Of_R_Package").

1. MICE – для заповнення пропусків даних

За наявності пропусків у вхідному наборі даних є пакет MICE для їх заповнення. Коли виникає проблема пропущених значень, найчастіше її рішення – прості заміни: нулями, середніми, модами і т. ін. Проте ні один з цих методів не адаптований до конкретних ситуацій і може призвести до невідповідності в даних. Пакет MICE замінює пропущені знання, використовуючи різноманітні техніки, в залежності від даних, з якими ви працюєте.

2. Rpart – класифікаційні та регресійні моделі

Пакет rpart в мові R використовується для побудови класифікаційних і регресійних моделей із застосуванням двокрокової процедури, а результат представляється у вигляді бінарних дерев. Найпростіший спосіб побудувати регресійне або класифікаційне дерево із застосуванням rpart – викликати функцію plot (). Сама функція plot () може не дати прийнятний результат, тому є альтернатива – prp () – потужна і гнучка функція. Функція rpart () дозволяє встановити відношення між залежною і незалежними змінними, щоб показати дисперсію залежної змінної на підставі незалежних. Наприклад, якщо компанія, що надає онлайн-навчання, хоче дізнатися, як на їх продажі (залежна змінна) впливає просування в соціальних мережах, газетах, реферативних посиланнях тощо, в rpart є кілька функцій, які можуть допомогти з аналізом цього явища.

rpart - абревіатура, яка розшифровується, як Recursive Partitioning and Regression Trees (рекурсивне розбиття і регресійні дерева). За допомогою rpart можна реалізувати як регресію, так і класифікаційні методи.

3. CARET: Classification And REgression Training

Пакет CARET – Classification And Regression Training (класифікація і регресійне навчання) – розроблений для комбінування моделей навчання і прогнозування. У пакеті є алгоритми, придатні для різних завдань. Спеціаліст з аналізу даних не завжди може точно сказати, який алгоритм кращий для вирішення того чи іншого завдання. Пакет CARET дозволяє підібрати оптимальні параметри для алгоритму за допомогою контрольованих експериментів. Метод перехресного пошуку, реалізований в цьому пакеті, шукає параметри, комбінуючи різні методи оцінювання продуктивності моделі. Після перебору всіх можливих комбінацій метод перехресного пошуку знаходить комбінацію, яка дає найкращі результати. Можна полегшити побудову моделей прогнозування завдяки спеціальним вбудованим функціям для розбиття даних, вибору важливих ознак, попередньої обробки даних, оцінювання важливості змінних, налаштування моделі через повторну вибірку та візуалізацію.

4. Пакет Nnet

Найбільш широко використовуваний і легкий для сприйняття пакет для роботи з нейронними мережами, але його обмеження – один рівень вузлів. Цей пакет не надає ніяких спеціальних методів для визначення кількості вузлів на прихованому рівні. Тому коли фахівці по роботі з великими даними застосовують nnet, завжди передбачається, що потрібно вказати значення, що лежить між кількістю вхідних і вихідних вузлів. Зокрема, нейронні мережі часто дають найкращі результати прогнозування із застосуванням функцій nnet, ніж стандартні методи прогнозування, такі як експоненціальне згладжування, регресія, тощо

Кожен пакет або функція в R має свої значення за замовчуванням. Перед тим як застосовувати алгоритм, має сенс дізнатися, які опції доступні. Значення за замовчуванням дадуть деякий результат, але немає впевненості, що він буде оптимальним або точним. У CRAN є й інші пакети для машинного навчання, наприклад, igraph, glmnet, gbm, tree, CORElearn, mboost та ін. Вони застосовуються в різних сферах для побудови найбільш адекватних моделей. Можна зіткнутися з ситуаціями, коли зміна одного параметра повністю поміняє вигляд вихідних даних.

Тому не варто покладатися на значення за замовчуванням: потрібно завчасно проаналізувати свої дані і вимоги до них та моделей перед застосуванням конкретного алгоритму.

1.6 Концепція системи підтримки прийняття рішень

На основі проведення порівняльного аналізу інструментів та моделей, передбачується розробка системи підтримки та прийняття рішень. Концепція розробки розглянутих вище та вже запропонованих підходів та архітектур необхідна для побудови математичних моделей оцінювання та аналізу фінансових ризиків.

Для розробки СППР з наведеною вище архітектурою використано такі системи: RStudio 1.1+та IDEWebStorm (Python, JavaScript). Інтерфейс користувача у вигляді web-сторінки інтуїтивно зрозумілий та створений таким чином, щобу користувача була можливість виконати всі етапи аналізу фінансових ризиків від моменту завантаження даних і до виведення остаточних результатів аналізу даних.

При створенні системи було використано пакет GNU, вихідний код для програмного середовища R написано на C, що є у вільному доступі за загальною публічною ліцензією GNU. Початкові дані для роботи (вхідні дані) можуть бути у різних текстових форматах: електронна таблиця Microsoft (Excel формату «*.xls» MSExcel 2003 або «*.xlsx» MSExcel 2007+), таблиця MSSQLServer та стандартні розширення файлів з даними Rstudio. Також існує можливість імпорту файлів з даними з іншого програмного забезпечення, наприклад таких систем: SPSS, SAS, Statata інших [40].

Головною вимогою до формату таблиці є те, що останньому стовпцеві повинна відповідати цільова змінна, яка може набувати значень: (1) 1 – «так»; (2) 0 – «ні». Оскільки весь масив вхідних даних спочатку перетворюється до текстового формату, то фізичний формат цільової змінної та інших змінних не має значення. При такому конвертуванні усіх даних у текстовий формат порожні

значення, значенняNULL та значення полів, які у текстовому форматі дорівнюють рядку, що у верхньому регістрі (uppercase) дорівнює «NULL», перетворюються у текстові «NULL»-значення, які надалі вважаються порожніми та завжди розміщуються в окрему категорію або інтервал [36, 40, 41].

Щодо дійсних чисел, то при подальшому виборі числового (numerical) формату змінної символи «крапки» та «коми» вважаються десятковими роздільниками. Також при завантаженні даних з MSExcel можна спочатку завантажити лише їх структуру: заголовки стовпців (headers) з кількістю рядків таблиці, а надалі скористатися комбінаціями клавіш копіювання та вставки власне даних лише з буферу обміну [40].

1.7 Висновки до розділу

Основними фінансовими ризиками, управління якими є ключовим фактором, що визначає ефективність фінансової діяльності – це кредитний та ринковий ризики. Діяльність банків та інших фінансових організацій здійснюється під впливом невизначеностей зовнішнього середовища (ринку, економіки, політики тощо), великої кількості змінних, контрагентів, осіб, поведінка яких не завжди може бути передбачена з прийнятною точністю. Відповідно виникає ризик, який безпосередньо пов'язаний з невизначеністю того чи іншого типу. Основне завдання з управління фінансовими ризиками зводиться до забезпечення стабільності у кожній сфері та налагодження зв’язків. Мінімізація фінансових ризиків здійснюється на двох рівнях – на рівні кожної окремої позики та на рівні кредитного портфеля в цілому. Для кожного рівня характерні свої особливості з менеджменту ризику.

В результаті виконання попереднього аналізу встановлено, що найпоширенішими методами оцінювання ризиків для поставленої задачі є лінійна і логістична регресії, дерева класифікації, нейронні мережі, мережа Байєса. Кожен з цих методів має свої переваги та недоліки.

Для підвищення ефективності прийняття об'єктивних рішень при аналізі кредитного та ринкового ризиків доцільно використовувати мережі Байєса та нечіткі нейронні мережі, які дають можливість враховувати невизначеності ймовірнісного та амплітудного типів. Ці підходи характеризуються швидкими алгоритмами навчання та нескладною інтерпретацією накопичених знань. Такі особливості обраних підходів роблять їх одними з найбільш перспективних і ефективних інструментів моделювання і оцінювання фінансових ризиків.

РОЗДІЛ 2
МОДЕЛЮВАННЯ КРЕДИТНИХ РИЗИКІВ

При наданні кредитів банк зобов’язаний дотримуватись основних принципів кредитування, зокрема проводити аналіз кредитоспроможності позичальників, дотримуватись встановлених банком вимог щодо концентрації ризиків. Одним із вагомих моментів під час видачі кредиту виступає кредитоспроможність та її загальна оцінка, а також гарант платоспроможності фізичної особи, який може бути представлений високим рівнем заробітної плати, платою поручителів або заставою особистого майна. Гарантія такого роду не завжди існує повною мірою для юридичних осіб, через що з'являються обхідні варіанти, а суми позик значно зростають.

Зазвичай на кредитоспроможність позичальника мають вплив різні фактори фінансового стану, це можуть бути власні кошти, ліквідні активи, та фінансова дисциплінованість. Існує два етапи за якими складається оцінка кредитоспроможності: врахування аналізу ділового ризику; аналіз фінансового стану позичальника для якого враховуються фінансові коефіцієнти і грошові потоки.

2.1 Застосування методів оцінювання інформативності змінних

При побудові моделей однією з труднощів є проблема відбору значущих для моделі вхідних ознак. Скорочення числа незалежних змінних необхідно для зменшення розмірності моделі не тільки з тим, щоб видалити з неї всі незначущі ознаки, що не несуть в собі корисної для аналізу інформації, і тим самим спростити модель, але і щоб усунути надлишкові ознаки. Дублювання інформації в складі надлишкової ознаки не просто не покращує якість моделі, але і часом, навпаки, погіршує її (як, наприклад, у випадку з мультиколінеарністю).

Очевидно, що одним з можливих виходів з ситуації, могла б стати побудова моделі на всіх можливих комбінаціях наборів вхідних ознак з

подальшим відбором того варіанту, який надавав би найкращу описову здатність результуючої ознаки *y* і при цьому містив мінімум незалежних змінних. Однак таке рішення можливе лише за наявності незначної кількості факторів-претендентів на включення у модель. У разі ж великого списку потенційних ознак подібна методика представляється досить складною, оскільки кількість моделей, які необхідно буде побудувати, виявляється досить великою. Зважаючи на це, необхідно застосовувати додаткові алгоритми відбору найбільш важливих факторів, які потребували б значно менших обчислень та витрат часу.

2.1.1 Обґрунтування вибору підмножини ознак

Метод відбору ознак намагається знайти підмножину вхідних змінних (які називаються ознаками або атрибутами). Є три стратегії – стратегія фільтра (наприклад, накопичення ознак), стратегія обгортання (наприклад, пошук згідно заданої точності) і стратегія вкладення (вибираються ознаки для додавання або видалення в міру побудови моделі, заснованої на похибках прогнозування).

Для застосування техніки відбору ознак використовуються такі підходи [42]:

− спрощення моделей, щоб зробити їх простішими для інтерпретації дослідниками / користувачами;

− зменшення часу на тренування (навчання) моделі;

− зменшення розмірності вхідних даних;

− покращене узагальнення шляхом скорочення перенавчання [30] (формально, зменшення дисперсії [42]).

Ідея використання техніки відбору ознак полягає у тому, що дані містять деякі ознаки, які є або зайвими, або значущими при формуванні результату, а тому можуть бути видалені без значних втрат інформації [43].

Техніки відбору ознак слід відрізняти від виділення ознак. Виділення ознак створює нові ознаки як функції від оригінальних ознак, в той час як відбір ознак повертає їх підмножину. Техніки відбору ознак часто використовуються у випадках, коли є багато атрибутів, а вибірки порівняно малі (мало точок даних).

Алгоритм відбору ознак можна розглядати як комбінацію технік пошуку для представлення нової підмножини ознак разом з обчисленням критеріїв, які відображають відмінність підмножини. Найпростішим алгоритмом є перевірка кожної можливої підмножини ознак і знаходження тієї, яка мінімізує величину похибки. Це вичерпний пошук по простору і він обчислювально важкий для наборів ознак, які не є дуже малими. Як показано нижче, вибір метрики сильно впливає на алгоритм і вони різні для трьох основних категорій алгоритмів відбору [44].

1. Методи обгортання використовують модель апріорної оцінки результату для оцінювання наборів ознак. Кожен новий набір використовується для тренування моделі, яка перевіряється на контрольній вибірці. На цій контрольній вибірці визначається число помилок (показник помилок моделі), яке надає оцінку для даної підмножини. Оскільки методи обгортання тренують модель для кожного набору, з обчислювальної точки зору вони дуже затратні, але дають, як правило, кращий набір ознак для конкретного типу моделі.

2. Метод фільтрів використовує непрямий показник замість показника помилки для оцінювання набору ознак. Цей показник вибирається так, щоб його можна було легко обчислити при збереженні показника корисності набору ознак. Зазвичай застосовуються такі показники: взаємна інформація [44], поточна взаємна інформація [45], коефіцієнт кореляції змішаних моментів Пірсона, алгоритм, заснований на показнику Relief [46] і відстань між класами / всередині класу або результат критеріїв значимості для кожної комбінації класів / ознак [45, 47]. Фільтри зазвичай обчислювально менш інтенсивні, але вони дають набори ознак, які не налаштовані на специфічний тип прогнозуючої моделі [48].

3. Методи вкладення є узагальнюючою групою технік, які здійснюють відбір ознак як частину процесу побудови моделі. Прикладом такого підходу є метод оцінювання коефіцієнтів лінійної регресійної моделі (Least absolute shrinkage and selection operator, LASSO) для побудови лінійної моделі, який штрафує коефіцієнт регресії на штраф L1, зменшуючи багато з них до нуля. Будь-які ознаки, які мають ненульові коефіцієнти регресії, «вибираються» алгоритмом LASSO [49].

2.1.2 Проведення загального статистичного тесту

На сьогоднішній день можна працювати з множиною числових показників, які аналізують та вимірюють ступінь взаємозв'язків між двома змінними – коефіцієнтами зв'язків. Першим із таких показників для оцінювання інформативності моделей є загальний статистичний тест та критерій x^2. Особливість цього підходу полягає у знаходженні залежностей між змінними – екзогенними та цільовою.

Цей критерій запропонував Карл Пірсон у 1990 році і використовував його для оцінювання близькості емпіричних розподілів до теоретичних. Критерій x^2 може використовуватися тоді, коли є необхідність встановити відповідність між двома порівнюваними рядами розподілу – емпіричним і теоретичним, або двох емпіричних. При цьому порівнюються частоти рядів розподілу, виявляються розбіжності між ними і визначається ймовірність цих розбіжностей [50].

За допомогою критерію виявляються відмінності в розподілах двох емпіричних рядів, порівнюються вибірки, які мають альтернативні ознаки, а також оцінюється ймовірність кореляції між альтернативними ознаками. Як і інші критерії згоди, представлені у роботі [51], критерій χ^2 являє собою деяку величину, що оцінюється з певною ймовірністю. Також за допомогою χ^2-критерію здійснюється статистична перевірка гіпотез відносно розподілів, тобто відповідність емпіричних даних розподілу деякому теоретичному закону розподілу. Така оцінка наближення емпіричного розподілу до теоретичного дає суму співвідношень частот. Збіг емпіричних і теоретичних частот зумовлює величину χ^2 =0. Це вказує на підтвердження нульової гіпотези, (H_0). При наявності достовірної різниці у частотах емпіричного і теоретичного ряду величина χ^2 буде свідчити про неправильність висунутої гіпотези [52].

Розподіл χ^2 сьогодні є одним з найбільш широко використовуваних у статистиці. Його використовують для перевірки статистичних гіпотез про ймовірний закон невідомого розподілу. Критерій також використовується для перевірки гіпотез стосовно різноманітних розподілів та розраховується за

формулою:

$$X^2 = \sum_{j=1}^{n} (\frac{f_{\in j} - f+}{f+})^2, \quad (2.1)$$

де $f_{\in j}$ – емпіричні значення;

$f+$ – теоретичні значення;

n – число ступенів свободи.

Для перевірки необхідно виконати порівняння емпіричних (спостережуваних) і теоретичних (вирахуваних у припущенні, наприклад нормального, розподілу) частот [53].

Якщо є повне узгодження емпіричних частот з частотами, вирахуваними або очікуваними, критерій χ^2 прирівнюють до нуля. В інакшому випадку, коли критерій не дорівнює нулю, це показує на невідповідність вирахуваних частот емпіричним частотам ряду. Для таких випадків необхідно оцінити значимість критерію χ^2, який теоретично може змінюватися у діапазоні від нуля до нескінченності. Це здійснюється шляхом порівняння фактично отриманої величини з її критичним значенням. Даний критерій дозволяє порівнювати розподіли частот незалежно від того, розподілені вони нормально чи ні. Зазвичай, із частотою виникнення події мають справу, коли змінні виміряні у шкалі найменувань та іншу характеристику, окрім частоти підібрати проблематично або неможливо. Іншими словами, коли змінна має якісні характеристики [54].

2.1.3 Рекурсивне виключення змінних

Для класифікації з невеликими навчальними вибірками та високою розмірністю вибір функції відіграє важливу роль для уникненні проблем із накладанням та покращенні результатів класифікації. Один з найбільш часто використовуваних методів відбору функцій для проблем з невеликими зразками – це метод рекурсивного усунення ознак (RFE). Метод RFE використовує можливість узагальнення, вбудований в алгоритми типу SVM (Support Vector

Machine), і тому підходить для задач з невеликими вибірками. Незважаючи на хороші показники, RFE прагне відкинути «слабкі» характеристики, що може забезпечити значне поліпшення продуктивності в поєднанні з іншими функціями.

Вибір критеріїв полягає у виборі підмножини відповідних функцій із більшого набору оригінальних визначених критеріїв, таких як ефективність класифікації або роздільність класу. Він відіграє значну роль у програмах машинного навчання. Вибір особливостей дуже важливий при невеликих проблемах класифікації вибірки, де кількість доступних навчальних зразків є дуже малою порівняно з кількістю ознак. Невелика класифікаційна вибірка є поширеною проблемою, що виникає у багатьох практичних задачах. Наприклад, при автоматичному розпізнаванні цілі отримують дані великих об’ємів від мульти/гіперспектральних датчиків [55]. Основна перевага вибору ознак у задачах класифікації на основі невеликих вибірок полягає у подоланні проблем із придатністю отриманих результатів для покращення показників прогнозування [56].

Серед різних методів вибору ознак рекурсивне усунення ознак (RFE) – нещодавно розроблений метод вибору ознак для проблем класифікації за невеликими вибірками [57]. RFE спочатку застосовується для класифікації кредитоспроможності, де об’єм навчальних даних є меншим 100, а кількість ознак - більше 10 і стає ефективним підходом у відборі ознак за невеликою вибіркою. Також даний підхід спрямований на покращення ефективності узагальнення, видаляючи найменш важливі функції, усунення яких матиме найменший вплив на помилки навчання. Крім того, RFE тісно пов'язаний з підтримуючими векторними машинами (SVM), які, як було показано, добре узагальнюють результати навіть в умовах класифікації на невеликих вибірках [56].

Незважаючи на те, що RFE-тест виявився перспективним для розв'язання задач при обробці невеликих вибірок, він прагне видалити зайві та слабкі функції та зберігає незалежні функції. Імовірно, що надлишкові функції можуть забезпечити краще розділення класів, і, наприклад, дві слабкі функції, які самі по собі не мають великого значення, можуть забезпечити значне поліпшення

результатів при спільному використанні. Таким чином, просте видалення зайвих або слабких елементів може погіршити ефективність класифікації.

У задачі лінійного розділення дискримінант функція має вигляд:

$$g(x_i) = w \times x_i + b, \tag{2.2}$$

де b – термін зміщення;

w – ваговий вектор;

x_i – навчальні дані, $x_i \in R^n, i = 1,...m$.

Для лінійно нероздільного випадку можна ввести зміщені змінні і ξ, які вимірюють відхилення точки даних від оптимальної гіперплощини:

$$\Phi(w,\xi) = \frac{1}{2} \| w \|^2 + C \sum_{i=1}^{n} \xi_i, \tag{2.3}$$

$$y_i (w \times x_i + b) \geq 1 - \xi_i, \xi_i \geq 0, \tag{2.4}$$

де, y, відповідає цільовій змінній:

$$y_i = \{\pm 1\}, i = 1,...m. \tag{2.5}$$

Задача оптимізації вирішується за рахунок застосування двоїстого підходу таким чином:

$$y_i (w \times x_i + b) \geq 1 - \xi_i, \xi_i \geq 0; \tag{2.6}$$

$$w(\alpha) = \sum_{i=1}^{m} \alpha_i - \frac{1}{2} \sum_{i,j=1}^{m} y_i y_j \alpha_i \alpha_j (x_i \times x_j); \tag{2.7}$$

$$0 \leq \alpha_i \leq C, i = 1,..., m; \tag{2.8}$$

$$\sum_{i=1}^{m} \alpha_i y_i = 0, \tag{2.9}$$

де α_i – коефіцієнти Лагранжа.

RFE тест зберігає найменш неважливі функції, які є найменш важливою ознакою, пов'язаною з найменшим значенням ваги в w. Для лінійних випадків ваговий вектор обчислюється так:

$$w = \sum_{k \in SV} y_k \alpha_k x_k. \tag{2.10}$$

Для нелінійних випадків обчислення виконуються за допомогою рівняння:

$$w_i = \frac{1}{2}\alpha^T K\alpha - \frac{1}{2}\alpha^T K(-i)\alpha, \qquad (2.11)$$

де $K(-i)$ вказує на обчислену матрицю ядра шляхом видалення i-ї функції на кожному вході x.

Підхід RFE рекурсивно видаляє функції на кожному кроці та повторно використовує решту елементів шляхом перенавчання на основі цих інших функцій. Якщо функція є слабкою, RFE її видалить. Однак слабка функція може все-таки бути важливим елементом при спільному використанні з іншими функціями. Таким чином, просто видалення зайвих або слабких ознак може погіршити ефективність класифікації. Для того щоб оцінити важливість потенційної слабкої ознаки, ефективність класифікації оцінюється після того, як ця ознака вперше буде видалена з точки зору відповідного значення. Якщо ефективність класифікації знижується після видалення цієї функції, ця функція зберігається, хоча вона має найменше значення. Цей процес повторюється до тих пір, поки не буде встановлено результат, що ефективність класифікації без цієї функції не буде погіршуватися [57].

2.1.4 Оцінка значущості результатів на основі дерев рішень

Одним з найбільш перспективних підходів для розв’язання задачі кредитоспроможності платників є дерево рішень – можливість представити правила в ієрархічній, послідовній структурі, де кожен об'єкт відповідає єдиному логічному користувачеві, що приймає рішення. Під правилом мають на увазі логічну конструкція, представлену у вигляді «якщо ... то ...».

На сьогодні існує значна кількість алгоритмів, які активно використовуються для розв’язання різних задач класифікації: CART, C4.5, NewId, ITrule, CHAID, CN2 та ін. [58].

Більшість із відомих алгоритмів відносять до так званих «жадібних алгоритмів». Якщо один раз було обрано атрибут і по ньому було створено розбиття на підмножини, то алгоритм «не може» повернутися назад і вибрати

інший атрибут, який був би більш значущим для отримання результату. Насправді, на етапі побудови дерева рішень неможливо сказати, чи отримаємо ми оптимальне розбиття.

Розглянемо алгоритм C4.5 побудови дерева рішень, для якого кількість потомків для вузла не обмежена. Цей алгоритм вирішує лише завдання класифікації, оскільки він «не вміє» працювати з постійним цільовим полем. Для вирішення поставленої задачі необхідно, по-перше, внести зміни в процедуру розбиття даних за знаннями постійного типу; а по-друге, що саме головне, ввести поняття "значущість" для вхідних атрибутів і визначити формулу для її розщеплення.

Алгоритм розбиття по значенню неперервного типу:

1. Упорядкувати всі дані (знання) по зростанню.

2. Розбити вихідну множину T на дві: T_1 і T_2. На першій ітерації в T_1 підпадає лише один первинний елемент, а всі інші – в T_2. На наступній ітерації первинний елемент з T_2 попадає в T_1 і т.д.

3. Обчислити індекс $Gini_{split}$ для кожної множини з *T*. Обрати ту множину, для якої індекс $Gini_{split}$ мінімальний. При цьому використовують такі варіанти обчислень:

$$Gini(T) = 1 - \sum_{i=1}^{n} p_i^2, \tag{2.12}$$

$$Gini_{split}(T) = \frac{N_1}{N} Gini(T_1) + \frac{N_2}{N} Gini(T_2), \tag{2.13}$$

де p_i – імовірність знаходження прикладу класу *i* у множині T;

N – кількість даних (значень) у множині T (N_1 та N_2) – у множині T_1 та T_2 відповідно.

4. Подальше розбиття вузла припиняється при виконанні однієї з умов:

- у вузлі міститься достатня кількість значень (параметр налаштування);
- вузол містить приклади одного класу;

– кількість нерозпізнаних прикладів менше мінімальної кількості прикладів у вузлі (параметр налаштування).

Тепер введемо поняття «значущості» вхідного атрибута. Під значимістю атрибута будемо розуміти показник, що характеризує, наскільки сильно вихідний результат залежить від вхідних показників.

Формула для розрахунку значущості має вигляд:

$$Z_m = \frac{\sum_{j=1}^{k_m}(E_{m,j} - \sum_{i=1}^{n_{m,j}} E_{m.j.i} \times \frac{N_{l,j,i}}{N_{l,j}})}{\sum_{l=1}^{g}\sum_{j=1}^{k_l}(E_{l,j} - \sum_{i=1}^{n_{l,j}} E_{l,j,i} \times \frac{N_{l,j,i}}{N_{l,j}})} \times 100\%, \quad (2.14)$$

де g – кількість вхідних атрибутів;

k_l – кількість вузлів, які було розбито за атрибутами *l*;

$E_{l,j}$ – ентропія батьківського вузла, розбитого за атрибутом *l;*

$E_{l,j,i}$ – ентропія дочірнього вузла для *j*-го елементу, розбитого за атрибутом *l*;

$N_{l,j,i}$, $N_{l,j}$ – кількість прикладів для відповідних вузлів;

$n_{l,j}$ – кількість дочірніх вузлів для *j*-го батьківського елементу.

Загалом, обчислення показників значущості для атрибутів можливо тільки після побудови дерева класифікаційних правил.

2.1.5 Остаточний метод включення змінних до моделі

Кожен із методів, який було розглянуто вище, надає оцінку впливу параметра на цільову змінну, через що потрібно враховувати оцінку за кожним з методів. Для цього проводиться скоринг за результатами оцінювання за кожним з методів, де для екзогенної змінної кожного з методів надається бінарний індикатор – чи варто вводити цю змінну в модель за версією конкретного методу:

$$\vec{\mu}_{i,j} = f_j(\vec{x}_i), \quad (2.15)$$

де $(\vec{x}_1, \vec{x}_2, ..., \vec{x}_n)$ – вектори екзогенних змінних;

n – кількість екзогенних змінних;

$(f_1, f_2, ..., f_m)$– функція методу, яка класифікує чи є змінна значущою;

n – кількість екзогенних змінних.

Таким чином, маємо кілька бінарних індикаторів для кожного використаного методу і для кожної із змінних. Далі обчислюється загальна оцінка для кожної змінної, яка визначає скільки методів обрали цю змінну як значущу для моделі. Загальна оцінка знаходиться в межах [0, n]. Якщо значення загальної оцінки більше або дорівнює $n/2$, змінна визнається значущою і автоматично додається в модель. На рисунку 2.1 показана блок-схема процесу автоматичного вибору змінних до моделі [58].

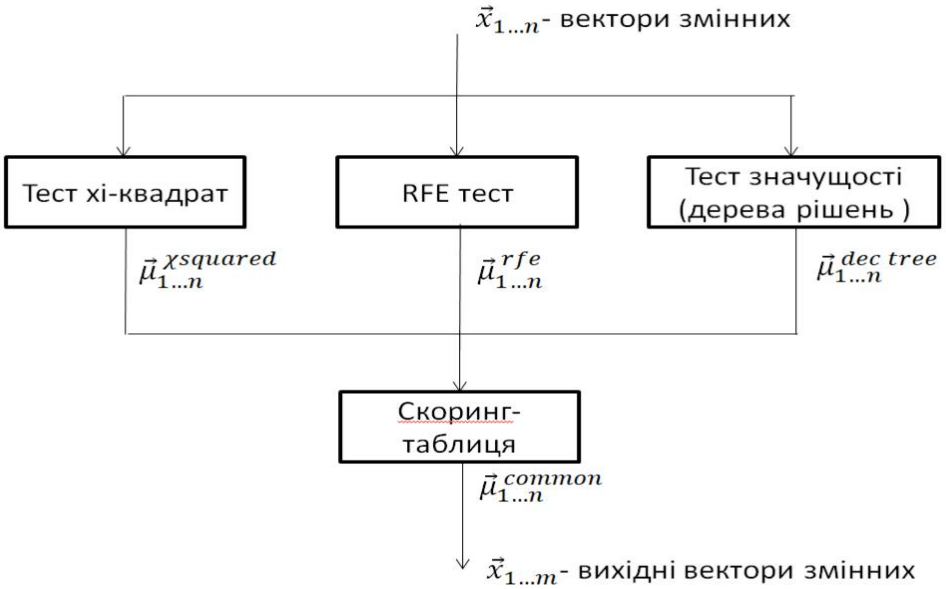

Рис. 2.1 Фрагмент схеми процесу остаточного вбору змінних

2.2 Оцінювання кредитного ризику на основі адаптивної мережі Байєса

При використанні байєсівських мереж (БМ) як інструменту для аналізу даних вирішуються одразу декілька математичних задач: (1) будується структура БМ, і (2) формується імовірнісний висновок. Задача побудови БМ за навчальними даними є NP-складною, тобто має нелінійну поліноміальну складність. Кількість всіх можливих нециклічних моделей, які потрібно проаналізувати, обчислюється за рекурентною формулою Робінсона,

запропонованою у 1976 році [59]:

$$f(n) = \sum_{i=1}^{n} (-1)^{i+1} \cdot C_n^i \cdot 2^{i\cdot(n-i)} \cdot f(n-i), \qquad (2.16)$$

де n – кількість вершин, а $f(0) = 1$. В реальності, при практичному виконанні повного перебору моделей, це можна зробити тільки для мереж не більш ніж з 7 вершинами (вузлами), тому що в іншому випадку для розв'язання задачі потребуються дуже великі обчислювальні ресурси.

Завдання формування остаточного результату застосування БМ – імовірнісного висновку, є важливим і складним і відноситься до класу задач прийняття рішень. Для реалізації цього етапу необхідно привести структуру БМ до так званого об'єднаного дерева (junction tree), а після цього використати алгоритм імовірнісного висновку на об'єднаному дереві, що ґрунтується на проходженні по дереву двох типів повідомлень: λ і π. Існуючі методи для побудови структури БМ, оцінювання її параметрів і формування висновку вимагають дуже трудомістких та витратних обчислень. Тому створення методів, що дозволяють зменшити обчислювальну складність відповідних процедур, є актуальною задачею при моделюванні процесів різної природи ймовірнісними мережами Байєса (або мережами довіри) [60].

Постановка задачі побудови байєсівської мережі (оцінювання структури і параметрів) складається з кроків, розглянутих нижче.

1. За евристичним методом побудова БМ складається із таких етапів: перший етап включає виконання обчислень значень взаємної інформації між усіма вершинами; на другому етапі виконується цілеспрямований пошук оптимальної структури, який використовує в якості оціночної функції оцінку мінімальної довжини (ОМД), яка застосовується і аналізується на кожній ітерації алгоритму навчання.

2. Далі ставиться завдання розробки методу формування (обчислення) імовірнісного висновку в БМ з реалізацією таких двох кроків; на першому кроці виконується обчислення матриці емпіричних значень спільного розподілу ймовірностей всієї мережі; на другому кроці виконується обчислення значень

ймовірностей для всіх можливих станів неінстанційованих вершин. Неінстанційованими вважаються ті вершини, які не отримують додаткової інформації крім апріорно відомої.

БМ являє пару $< G, B >$, в якій перша компонента G – це спрямований ациклічний граф, вершини якого відповідають випадковим змінним, він формально записується у вигляді множини умов незалежності: кожна змінна незалежна від її батьків в G. Друга компонента пари B – це множина параметрів, що визначають мережу; ця множина містить параметри $\Theta_{X^{(i)}|pa(X^{(i)})} = P(X^{(i)}|pa(X^{(i)}))$ для кожного можливого значення $x^{(i)} \in X^{(i)}$ і $pa(X^{(i)}) \in Pa(X^{(i)})$, де $Pa(X^{(i)})$ позначає множину батьків змінної $X^{(i)} \in G$. Кожна змінна $X^{(i)} \in G$ представляється у вигляді вершини. Якщо для формального представлення задачі розглядають більше одного графа, то для визначення батьків деякої змінної $X^{(i)}$ на графі використовують позначення $Pa^G(X^{(i)})$. Повна спільна ймовірність (спільний розподіл) для змінних БМ обчислюється за виразом [61]:

$$P_B(X^{(1)},...X^{(N)}) = \prod_{i=1}^{N} P_B(X^{(i)} \mid P_a(X^{(i)})). \qquad (2.17)$$

Гібридна мережа Байєса $B = (X, D, P)$ визначається через спрямований ациклічний граф $G = (X, E)$ і його функції $P_i = \{P(x_i \mid pa_i)\}$, де pa_i – множина батьківських вузлів x_i. X – множина змінних, розділених на дискретні Δ і неперервні Γ змінні, тобто $X = \Gamma \cup \Delta$. Структура графа G є обмеженою, через що неперервні змінні не можуть мати дискретне представлення як і їх вузли-нащадки. Тоді умовний розподіл неперервних змінних буде задаватися за лінійною гаусівською моделлю:

$$P(x_i \mid I = i, Z = z) = N(\alpha(\alpha(+ \beta(i) \times z, \gamma(i)), \quad x_i \in \Gamma, \qquad (2.18)$$

де Z та I – множини відповідно неперервних і дискретних батьків x_i, $N(\mu(\sigma)$ – мультиноміальний нормальний розподіл. Мережа представляє собою спільний

розподіл усіх його змінних, заданих добутком усіх таблиць умовних ймовірностей (ТУЙ), які містять параметри моделі. Тобто параметрами БМ є умовні ймовірності у відповідних таблицях.

Формулювання ймовірнісного висновку за мережею можна розглядати як процедуру перетворення багатовимірних гаусіанів у мережу Байєса. Змінні упорядковуються у певному порядку $X_1,\ldots,X_n$. Потім цей висновок використовується для знаходження умовного розподілу:

$$P(X_i \mid X_1,\ldots,X_{i-1}) = N(X_i; \beta_{i,0} + \sum_{j=1}^{i-1} \beta_{i,j} X_j, \sigma_i^2). \quad (2.19)$$

Ребро з X_j до X_i $(1 \le j < i)$ створюється тоді і тільки тоді, коли $\beta_{i,j} \neq 0$. Умовний ймовірнісний розподіл (УЙР) X_i називають спільним розподілом, що має вигляд деякого виразу після скорочення усіх нульових значень. Лінійний умовний ймовірнісний розподіл для кореневих вузлів є просто одновимірною гаусіаною. Мережа Байєса, у якій всі умовні ймовірнісні розподіли є лінійними, називається багатовимірною лінійною гаусіаною (ЛГ) [61]. Таким чином, у даному випадку, кожна багатовимірна гаусіана може бути представлена лінійною гаусіаною. Зворотне твердження також є справедливим. Кожна МБ з лінійними УЙР являє собою спільний нормальний розподіл.

Умовна лінійна гаусіана (УЛГ) – це МБ, яка містить як неперервні змінні (Γ), так і дискретні змінні (Δ), з такими обмеженнями [62]:

− дискретний вузол не може мати неперервних батьків; таким чином, всі УЙР для дискретних вузлів можуть бути подані так само як у дискретних мережах Байєса;

− УЙР будь-якої неперервної змінної є лінійним УЙР, заданий будь-якою комбінацією дискретних батьків. Формально, якщо вузол Y має батьків $\{X_1,\ldots,X_k\} \subseteq \Gamma$ і $D = \{D_1,\ldots,D_l\} \subseteq \Delta$, він визначається як УЙР з використанням таких параметрів: для кожного $d \in Dom(D)$, $\beta_{d,0},\ldots,\beta_{d,k}$ і σ_d^2:

$$P(Y \mid x,d) = N(Y; \beta_{d,0} + \sum_{i=1}^{k} \beta_{d,i} x_i, \sigma_d^2). \quad (2.20)$$

Це найбільш популярний вид гібридних мережевих моделей. Дані моделі

дають можливість відтворити тільки лінійні відношення між неперервними змінними та не передбачують те, що дискретні змінні мають неперервних батьків. Завдяки математичній зручності дані моделі отримали широке застосування.

В УЛГ задаються значення дискретних змінних та їх розподіл у вигляді багатовимірної гаусіани. Спільний розподіл представлений у вигляді композиції гаусіан, з якими можна працювати застосовуючи аналітичні інструменти. При умові, коли всі дискретні змінні буде задано, умовні ймовірнісні розподіли неперервних змінних представляють собою лінійні умовні неперервні розподіли. Лінійна гаусіана та умовна лінійна гаусіана представляють нормальний розподіл при наданні будь-яких значень дискретним змінним. Через це слідує, що спільний розподіл, представлений умовною лінійною гаусіаною, буде композицією цих гаусіан, кожна з яких відповідає реалізації дискретних змінних.

Таким чином, використання МБ дає можливість аналізувати причинно-наслідкові зв'язки між окремими змінними (подіями, даними) і формулювати на цій основі обґрунтований ймовірнісний висновок (прогноз). Використання гібридних МБ дає можливість коректно аналізувати неперервні змінні у моделі (наприклад, вік чи сума кредиту) і досягати вищої точності остаточного результату [63].

Розглянемо реалізацію процедури прогнозування кредитоспроможності індивідуального позичальника за допомогою Байєсівських мереж. Нехай за рік банком видано 1600 кредитів. Кожний клієнт описується 18 характеристиками (вік, сума кредиту, сімейний стан, рівень заробітної плати, наявність заборгованості за іншими кредитами, і т. ін.). За наявними даними будується мережа, яка показує зв'язок між характеристиками клієнта і вершиною – подією повернення кредиту. За навчальними даними визначається ймовірність повернення кредиту новим клієнтом, що вже прийшов до банку. Отже, треба визначити PD_i – ймовірність дефолту потенційного позичальника. Оскільки числові характеристики (сума кредиту, вік, дохід і т.ін.) набувають багато

значень, тобто є неперервними, то для розв'язання даної задачі необхідно використовувати гібридні мережі Байєса [2]. За описаною вище методикою необхідно виконати докладний аналіз проблеми і зробити формалізовану постановку задачі. Розглянемо детальніше схему застосування запропонованої методики.

1 крок. Для розв'язання задачі оцінювання кредитоспроможності позичальника необхідно зібрати необхідні статистичні дані за виданими кредитами, частина з яких була повернута, а частина виявилась дефолтами, тобто отримати позитивні і негативні приклади для навчальної бази даних. На основі вибраних параметрів позичальника та кредиту необхідно розробити формальну модель і оцінити ймовірності дефолту позичальника PD_i :

$$PD_i = F(w^j, x_i^j), \tag{2.21}$$

де w^j – ваги параметрів x_i^j , i – кількість позичальників, j – кількість параметрів кредиту. Модель для оцінювання кредитоспроможності на основі мереж Байєса описується таким чином:

$$PD_i = F(v_i^k, G, J) = 1 - PR_i, \tag{2.22}$$

де v_i^k – змінні, що описують характеристики клієнта і кредиту; J – імовірнісний розподіл змінних v_i^k; G – спрямований ациклічний граф, вузли якого відповідають випадковим змінним v_i^k модельованого процесу; PR_i – ймовірність повернення кредиту.

2 крок. Дані, що необхідні для розв’язання задачі, це початкові статистичні дані за N виданими кредитами, з яких приблизно 20% випадків дефолтів, а 80% – повернутих кредитів.

3 крок. Показники та характеристики, що описують клієнта та його стан, для цієї задачі є взаємовиключними змінними.

4 – 5 кроки. На даних кроках для неперервних змінних використовується дискретизація (коли область значень змінної поділяється на рівні або проміжки). Ця процедура виконується через те, що в задачі використовуються неприривні

змінні, а також будується гібридна мережа Байєса. Користувач визначає кількість проміжків особисто. У випадку, коли для стовпчиків не було вказано кількість проміжків для розбиття, ці стовпчики автоматично поділяються на два інтервали.

Доцільним є виконання дискретизації даних за однаковою шириною класів або, в залежності від інтерпретації, за однаковою кількістю точок всередині кластерів. При такому підході ширина та кількість таких інтервалів регламентується банком, базуючись на соціологічних або демографічних дослідженнях кожної групи клієнтів [63]. При побудові структури мережі Байєса слід пам'ятати, що обраний алгоритм впливає на швидкодію виконання програми і на вигляд самої побудованої структури.

2.3 Застосування регресійних підходів до моделювання кредитних ризиків

2.3.1 Застосування лінійних ймовірнісних моделей

Лінійна ймовірнісна модель представляє собою регресійну модель, в якій значення залежної змінна дорівнює 0 або 1 залежно від того, чи було затверджено (прийнято) конкретну заяву, чи ні. З математичної точки зору, правило прийняття рішення виражається в такий спосіб:

$$y = \beta_0 + \beta_1 x_1 + ... + \beta_k x_k + \varepsilon, \tag{2.23}$$

де y – залежна цільова змінна;

$(x_1, x_2, ..., x_k)$ – незалежні змінні;

$(b_1, b_2, ..., b_k)$ – вагові коефіцієнти (ваги), присвоєні незалежним змінним;

ε – це випадкова похибка моделі, зумовлена неточністю її структури, оцінок параметрів, похибками вимірів та обчислювальних процедур.

Розподіл випадкової похибки залежить від типів розподілів незалежних змінних. При цьому її математичне очікування має дорівнювати нулю.

У векторному представленні має вигляд:

$$y = b'x + \varepsilon, \tag{2.24}$$

де x – вектор пояснюючих змінних;

$b^{'}$ – транспонований вектор параметрів.

Отже,

$$P(y \mid x) = b^{'}x. \tag{2.25}$$

Умовну ймовірність можна також інтерпретувати як ймовірність схвалення заявки, що відноситься до групи параметрів *x*. Оцінена ймовірність затвердження заявки може бути витлумачена аналогічним чином. Через це маємо задачу регресійного оцінювання вагових коефіцієнтів і розрахунку ймовірності затвердження нової заявки. Коли приймається рішення про надання кредиту, отриманий таким чином результат слід порівняти з межею (порогом) відсічення. Варто зазначити, що така модель досить проста в реалізації, але не завжди дає прийнятний результат прогнозування. Отже, задача полягає у необхідності оцінити вектор параметрів моделі $(\beta_1, \beta_2, ..., \beta_k)$ на основі наявних експериментальних значень y і $(x_1, x_2, ..., x_k)$; потім застосувати модель для оцінювання значення залежної змінної для потенційного клієнта.

2.3.2 Моделювання на основі логістичної регресії

Модель логістичної регресії представляє собою різновид нелінійної множинної регресії для проведення аналізу на функціональну залежність між декількома регресорами (які представляють незалежні змінні) та між залежною змінною. Використання бінарної логістичної регресії доцільно в тому випадку, коли на вхід моделі подається тільки два значення. Через це бінарна логістична регресія частіше за все використовується в задачах оцінки кредитоспроможності. Саме в цих задачах дотримується правило для двох значень щодо повернення кредиту, які подаються на вхід: 0 - у разі виплати кредиту, 1 - у разі невиплати кредиту. Також ці значення можуть бути представлені у вигляді True/False або Non_default/ Default. У загальному вигляді регресійна модель буде наступною:

$$y = F(x_1, x_2, ..., x_n) \tag{2.26}$$

Також допускається, що лінійна функція незалежних змінних є залежною (тільки у випадку множинної регресії):

$$y = b_1 x_1 + b_2 x_2 + ... + b_n x_n + u, \quad (2.27)$$

де у – залежна змінна (результат прийняття рішення);

x_i – пояснювальна змінна (критерій);

b_i – ваги пояснювальної змінної;

u – випадкова похибка.

У векторному вигляді цей вираз виглядає так:

$y = b'x + u,$

де x – вектор пояснювальних змінних;

b' – транспонований вектор параметрів пояснювальних змінних. В такому випадку умовна ймовірність події обчислюється за допомогою виразу

$$P(y \mid x) = b'x \quad (2.28)$$

При застосуванні бінарної логістичної регресії ймовірність входження позичальника у стан дефолту D пов’язана із змінними $(X_1, X_2, ..., X_k)$. Логістична регресія являє собою один із різновидів множинної регресії. Її загальне призначення полягає в аналізі наявних зв'язків між декількома незалежними змінними (регресорами або предикторами) і залежною змінною. Бінарна логістична регресія може застосовуватися тільки у тому випадку, коли залежна змінна є бінарною (приймає тільки два значення). Іншими словами, за допомогою логістичної регресії можна оцінювати ймовірність того, що подія відноситься до конкретного типу випробування (хворий/здоровий, повернення кредиту/дефолт і т. ін.). Це може досягатися із застосуванням такого регресійного рівняння (логіт-перетворення) та бути представлено за формулою [21, 22]:

$$P = \frac{1}{1 + e^{-y}}, \quad (2.29)$$

де Р – ймовірність того, що відбудеться подія, яка цікавить;

е – основа натуральних логарифмів 2,71;

y – змінна, що визначається рівнянням лінійної регресії.

Залежність, що зв'язує ймовірність події і величину *y*, показана на рисунку 2.1.

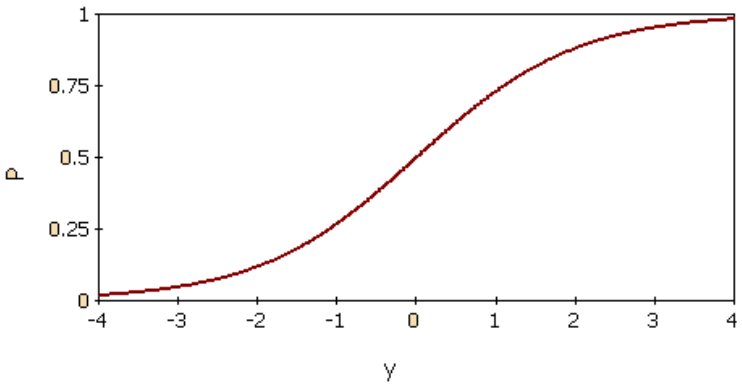

Рис. 2.2 Логістична функція розподілу

Логістичний зв’язок подано у наступним рівнянням:

$$\ln(\frac{p}{1-p}) = \beta_0 + \beta_1 X_1 + ... + \beta_n X_n, \tag{2.30}$$

де оцінка β_0 – перетин;

оцінки β_i – вагові коефіцієнти для незалежних змінних.

Перевага цієї моделі полягає в тому, що правильно визначається ймовірність між значеннями нуля та одиниці незалежно від характеру значень, що подаються на вхід. Ймовірність невиплати кредиту записується у більш зручній формі завдяки векторам для зображення параметрів моделі та має наступний вид:

$$p = \frac{\exp(\beta X)}{1+\exp(\beta X)}. \tag{2.31}$$

Бувають випадки, коли застосування множинної регресії для визначення ймовірності повернення чи неповернення позичальником кредиту дає результат, який виходить за межі значень інтервалу [0; 1]. Це стає через те, що бінарна модель не має інформації про вхідну змінну та те, що ця змінна є бінарною. В такому випадку виконується трансформація виразів, що описані вище, та які задовольняють вимогам стосовно знаходження вихідної змінної в інтервалі [0;

1]. Ймовірність визначається із припущенням про нормальний розподіл та записується як:

$$p = \Phi(b'x) = \int_{-\infty}^{b'x} \varphi(z)dz, \tag{2.32}$$

де $\Phi(z)$ – щільність нормального розподілу.

Відповідна трансформація призводить до використання пробіт-моделі.

У випадку, якщо обирається логістична функція розподілу для подання ймовірності *p*, то даний перехід веде до використання логіт-моделі. У цьому випадку ми отримуємо [2, 3]:

$$p = \Phi(b'x) = \int_{-\infty}^{b'x} \varphi(z)dz = \frac{1}{1+e^{-b'x}} \tag{2.33}$$

або в іншому вигляді:

$$p = \frac{e^{b_1x_1+\ldots+b_nx_n}}{1+e^{b_1x_1+\ldots+b_nx_n}}. \tag{2.34}$$

У порівнянні з функцією нормального розподілу, модель логістичної функції розподілу має закриту форму. Це може призвести до того, що обчислення в логіт-моделі буде набагато простішими ніж обчислення у пробіт-моделі. Для визначення значень вагових коефіцієнтів b_i використовується метод максимальної правдоподібності.

2.4 Оцінювання кредитоспроможності позичальника за допомогою нечітких нейронних мереж

Для врахування якісних і кількісних характеристик позичальника, вектор даних представляється за допомогою лінгвістичної змінної, яка визначається як <b,T,X,G,M>, де:

- b – ім’я лінгвістичної змінної;
- T – є базовою терм-множиною лінгвістичної змінної, кожний елемент якої (терм) є нечіткою множиною на універсальній множині X;

– G – є синтаксичним правилом, часто представленим у вигляді граматики, що дає можливість оперувати елементами терм-множини T, генерувати нові терми;

– M – семантичне правило, що задає функцію належності (ФН) нечітких термів, утворених синтаксичними правилами [63]. Для побудови правила необхідно побудувати функції належності кожної лінгвістичної змінної, тобто задати вид функції та її параметри. Параметри функцій належності знаходять в процесі її побудови. Підхід на основі нечітких нейронних мереж дозволяє автоматично будувати функцію належності, а відповідно й створювати базу знань та реалізувати нечіткий логічний висновок. В загальному вигляді алгоритм має етапи, подані нижче.

1. Визначення множини вхідних змінних про позичальника: $\vec{X} = \{X_1, X_2, ..., X_i, ..., X_N\}$.

2. Визначення множини вихідних змінних – можливі групи ризику: $D = \{D_1, D_2, ..., D_i, ..., D_M\}$.

3. Формування базової терм-множини з відповідними функціями належності до кожної змінної: $A = \{a_1, a_2, ..., a_N\}$.

4. Формування кінцевої множини нечітких правил, які узгоджуються до змінних, які в них використовуються.

5. Реалізація логічного висновку, іншими словами визначення істинності для передумов кожного правила та визначення нечітких підмножин для змінних виходу для кожного правила.

6. Композиція нечітких підмножин кожної змінної виходу за всіма правилами.

7. Знаходження чіткого значення для кожної з вихідних лінгвістичних змінних.

2.4.1 Оцінювання кредитоспроможності позичальника за допомогою ННМ з логічним висновком Мамдані

В нечіткій нейронній мережі (ННМ) з логічним висновком Мамдані реалізується логічний висновок, який має етапи, подані нижче [64].

1. Введення нечіткості. Знаходиться ступінь істинності для передумов кожного правила: $A_1(x_0)$, $A_2(x_0)$, $B_1(x_0)$, $B_2(x_0)$.

2. Логічний висновок. Знаходимо рівні «відсікання» для передумов кожного з правил (з використанням операції знаходження мінімуму): $\alpha_1 = A_1(x_0) \cap B_1(y_0)$; $\alpha_2 = A_2(x_0) \cap B_2(y_0)$. Знаходимо рівні «відсікання» функції належності: $C_1' = (\alpha_1 \cap C_1(z))$; $C_2' = (\alpha_2 \cap C_2(z))$.

3. Композиція. Знаходимо об'єднання знайдених відсічених функцій належності з використанням операції максимум та отримуємо підсумкову нечітку підмножину для змінної виходу з функцією належності: $\mu_\Sigma = C(z) = C_1'(z) \cup C_2'(z) = (\alpha_1 \cap C_1(z)) \cup (\alpha_2 \cap C_2(z))$.

4. Зведення до чіткості з використанням, наприклад, центроїдного методу.

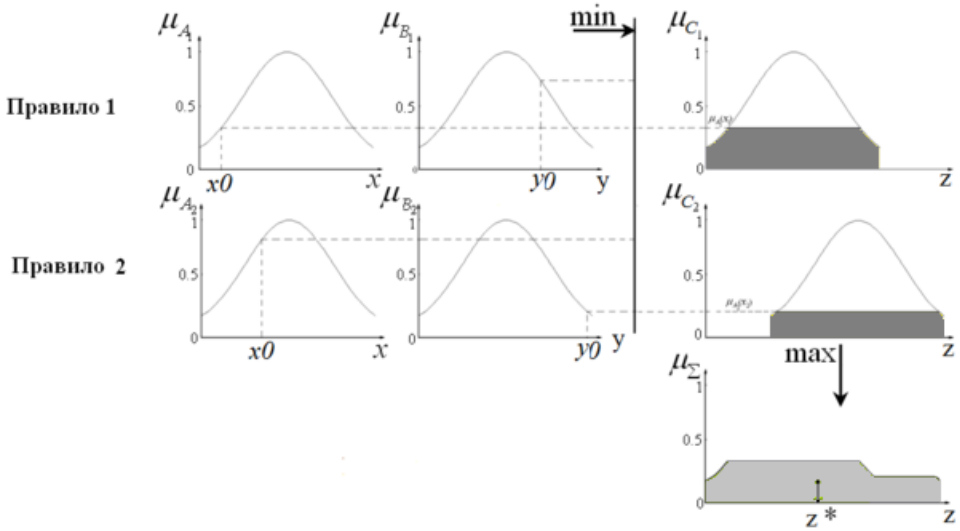

Рис. 2.3 Ілюстрація логічного висновку Мамдані

Вхідними є такі фактори впливу, як вік, рівень заробітної плати, строк кредитування, а вихідним – ймовірність повернення кредиту та відповідний фінансовий клас позичальника. Задачу оцінки кредитоспроможності можна сформулювати таким чином: Кожна кредитна заявка задається вектором $\{x_1, x_2, .., x_i, .., x_n\}$, де x_i – певним чином формалізовані дані з анкети позичальника та параметри кредиту. Далі, за заданим вектором потрібно прийняти рішення про надання кредиту, тобто оцінити ймовірність повернення кредиту [63, 64]. В

загальному вигляді нечіткий логічний висновок має такі етапи: визначення множини вхідних змінних; визначення множини вихідних змінних; формування базової терм-множини з відповідними функціями належності кожного терма; формування кінцевої множини нечітких правил, узгоджених щодо використовуваних в них змінних. Знаходження чіткого значення для кожної з вихідних лінгвістичних змінних. Точність моделі та кількість помилок І-го та ІІ-го роду залежать від порогу відсікання, який буде встановлений організацією. Встановлення порогу відсікання дозволяє визначити не тільки процент клієнтів, який було відсіяно, а також нижню границю ймовірності повернення кредиту (нижнє значення для клієнта, коли він не зможе повернути кредит). За допомогою моделей на основі нечіткої логіки є можливість оцінювати кредитоспроможність позичальників та класифікувати їх як надійних або ненадійних за рахунок чого зменшувати кількість клієнтів, що були відсіяні [64].

2.4.2 Оцінювання кредитоспроможності позичальника за допомогою ННМ з логічним висновком Сугено

Поєднання нейронних мереж з нечіткою логікою привело до появи множини нечітких нейронних мереж: мережа TSK (Takagi, Sugeno, Kang'a), що базується на логічному висновку Сугено; мережа ANFIS, що також базується на логічному висновку Сугено та інші [64]. Для розв'язання задачі оцінювання кредитоспроможності позичальника можливо використовувати мережу TSK з висновком Сугено. Проілюструємо нечіткий висновок Сугено. Нехай логічний висновок спирається на таку базу правил:

$$П_1\text{: якщо } x \in A_1 \text{ і } y \in B_1\text{, то } z=a_1x+b_1y,$$

$$П_2\text{: якщо } x \in A_2 \text{ і } y \in B_2\text{, то } z=a_2x+b_2y,$$

де x і y вхідні змінні, z – вихідна змінна, A_1, A_2, B_1, B_2, C_1, C_2 – деякі задані функції належності, a_1, a_2, b_1, b_2 – деякі числа [65].

Обчислювальний алгоритм має такий вигляд:

1. Введення нечіткості як в алгоритмі Мамдані.

2. Обчислення нечіткого висновку; знаходимо

$\alpha_1 = A_1(x_0) \cap B_1(y_0), \alpha_2 = A_2(x_0) \cap B_2(y_0)$,та індивідуальні виходи правил за формулами, представленими нижче:

$$\dot{z}_1 = a_1 x_0 + b_1 y_0, \tag{2.35}$$

$$\dot{z}_2 = a_2 x_0 + b_2 y_0. \tag{2.36}$$

3. Визначення чіткого значення змінної виходу:

$$z_0 = \frac{\alpha_1 \dot{z}_1 + \alpha_2 \dot{z}_2}{\alpha_1 + \alpha_2} \tag{2.37}$$

На рисунку 2.4 подана ілюстрація алгоритму обчислення нечіткого висновку Сугено:

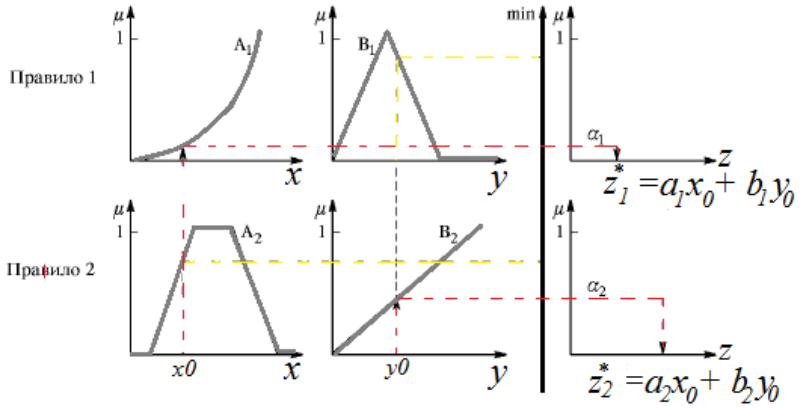

Рис. 2.4 Нечіткий логічний висновок Сугено

Для більш детального та наочного прикладу правила мережі представимо у такому вигляді [19]:

$$R_1: \text{якщо } x_1 \in A_1^{(1)}; x_2 \in A_2^{(1)}; \ldots, x_n \in A_n^{(1)}; \text{ то } y_1 = p_{10} + \sum_{j=1}^{N} p_{1j} x_{j;}$$

$$R_M: \text{якщо } x_1 \in A_1^{(M)}; x_2 \in A_2^{(M)}; \ldots, x_n \in A_n^{(M)}; \text{ то } y_M = p_{M0} + \sum_{j=1}^{N} p_{Mj} x_{j;}$$

де $A_i^{(k)}$ – значення лінгвістичної змінної x_i для правила R_k з функцією належності

$$\mu_A^{(k)}(x_i) = \frac{1}{1 + (\frac{x_i - c_i^{(k)}}{\sigma_i^{(k)}})^{2b_i^{(k)}}}. \quad (2.38)$$

Композиція результатів має вигляд:

$$y(k) = \frac{\sum_{k=1}^{M} w_k y_k(x)}{\sum_{k=1}^{M} w_k}, \quad (2.39)$$

де $\mu_A^{(k)}(x)$– ступінь виконання умов правила

$$\mu_A^{(k)}(x) = \prod_{j=1}^{N} \left[\frac{1}{1 + (\frac{x_i - c_i^{(k)}}{\sigma_i^{(k)}})^{2b_i^{(k)}}} \right]. \quad (2.40)$$

В ННМ з висновком Сугено виділяють 5 шарів. Опишемо, що відбувається в кожному шарі мережі [65]:

1. Перший шар виконує роздільну фазифікацію кожної змінної x_i i=1,2…N, визначаючи для кожного k-го правила висновку значення функції належності $\mu_{A^{(\kappa)}}(x_i)$ відповідно до функції фазифікації, наприклад дзвіноподібної. Це параметричний шар з параметрами $c_j^{(k)}$, $\sigma_j^{(k)}$, $b_j^{(k)}$, які мають бути налаштовані в процесі навчання мережі.

2. Другий шар виконує агрегування окремих змінних x_i, визначаючи результуючу ступінь належності w_k вектора X умова k-го правила, де

$$w_k = \mu_A^{(k)}(x) = \prod_{j=1}^{N} \left[\frac{1}{1 + (\frac{x_i - c_i^{(k)}}{\sigma_i^{(k)}})^{2b_i^{(k)}}} \right] \quad (2.41)$$

3. В третьому шарі розраховується значення $y_k(x) = p_{k0} + \sum_{j=1} p_{kj} x_j$, а також виконується множення функцій $y_k(x)$ і w_k, що були сформовані у попередньому шарі. Це параметричний шар з параметрами p_{k0}, p_{kj}, які мають

бути налаштовані.

4. В четвертому шарі розраховується зважена сума сигналів $y_k(x)$ і сума ваг

$$\sum_{k=1}^{M} w_k . \tag{2.42}$$

5. У п'ятому шарі ваги мають бути нормалізовані і розраховується вихідний сигнал мережі таким чином:

$$y(x) = \frac{\sum_{k=1}^{M} w_k y_k(x)}{\sum_{k=1}^{M} w_k}. \tag{2.43}$$

Для збільшення точності висновку також використовується рекурентна нечітка нейронна мережа TSK, у якій виходи кожного правила подаються знову на вхід мережі разом з вхідними змінними x_i . Таким чином, запам'ятовується минуле значення виходу кожного правила, яке використовується для обчислення нового висновку.

2.5 Висновки до розділу

Процес аналізу кредитоспроможності позичальників кредитів включає в себе розробку методів та критеріїв аналізу процесу кредитування; оцінку потенційного клієнта, а також супровід позики після видачі кредиту. Банкам та організаціям для оптимізації кредитного процесу необхідно знаходити компроміс між якістю та ефективністю всього процесу кредитування. Оцінка кредитного ризику здійснюється за допомогою оцінки кредитоспроможності позичальника. Найбільш поширеним методом оцінювання кредитоспроможності позичальника є скорингові карти, тобто набір характеристик та балів, що їм привласнюються у процесі оцінювання клієнта.

Також зручним і прозорим формальним представленням є опис позичальника за допомогою лінгвістичних змінних. Для обчислення нечіткого висновку постає задача побудови нечіткої бази знань. Для цього необхідно

описати відповідні функції належності. Зазвичай такий опис проводиться за допомогою експертів, що вносить суб’єктивність в процедуру оцінювання кредитного ризику. В такому випадку найкраще використовувати ННМ в яких результат отримуються на основі нечітких логічних висновків, а параметри функцій належності налаштовуються за допомогою алгоритмів нейронних мереж. Пріоритетною являється задача мінімізації ризику, через що відразу відкидаються ненадійні позичальники, а для всіх інших розв’язується задача максимізації доходу портфеля позик.

РОЗДІЛ 3
МОДЕЛЮВАННЯ РИНКОВИХ РИЗИКІВ

В сучасних умовах загальної економічної нестабільності, коли світ входить у період фінансово-економічної кризи, з'являється необхідність в удосконаленні та подальшому ефективному розвитку світової фінансової системи. Одним із найважливіших напрямів є, на сьогодні, виконання поглиблених наукових досліджень в напрямах, пов'язаних з математичним моделюванням і прогнозуванням нелінійних нестаціонарних процесів, характерних для фінансової сфери.

Більша частина досліджуваних процесів в сучасній економіці, фінансах, екології, соціально-економічних дослідженнях є сьогодні нелінійними і нестаціонарними або частково (на окремих часових інтервалах) лінійними і стаціонарними. Такі процеси, зазвичай, характеризуються наявністю стохастичних або детермінованих трендів та напряму залежать від конкретних випадкових факторів впливу і стохастичних збурень. Значна кількість досліджуваних процесів у фінансах є гетероскедастичними або інтегрованими, тобто їх умовна дисперсія або математичне очікування змінюються в часі на інтервалі дослідження. Як правило, нестаціонарні процеси проявляють нелінійність відносно параметрів або змінних, а тому їх називають нелінійними нестаціонарними процесами. Моделі гетероскедастичних процесів включають рівняння, які описують амплітуду процесу (сам процес), і рівняння, що описують динаміку його дисперсії.

3.1 Тестування наявності гетероскедастичності

Гетероскедастичність (heteroscedasticity) – поняття, яке використовується в галузі прикладної статистики та означає неоднорідність спостережень, що виражається у змінній дисперсії випадкових залишків (похибок) регресійних моделей. Фактично, ці залишки визначають характер процесу. Наявність

гетероскедастичності випадкових похибок може призвести до неефективності оцінок, які було отримано за допомогою методу найменших квадратів або інших методів оцінювання параметрів математичних моделей [67]. Крім того, даний випадок призводить до зміщення оцінок і неспроможності виконання класичного аналізу коваріаційної матриці МНК-оцінок параметрів моделі. Виходячи з цього маємо, що статистичні висновки про якість отриманих оцінок можуть бути неадекватними. У зв'язку з цим тестування моделей на гетероскедастичність є однією з необхідних процедур при створенні математичних моделей на основі статистичних даних [67].

Дуже часто виявлення проблеми гетероскедастичності можна передбачити завчасно, засновуючись на аналізі наявних статистичних даних та їх характеру. Такі випадки дозволяють застосовувати відповідні заходи щодо усунення цього ефекту на етапі специфікації моделі регресії що дозволяє зменшити, або взагалі усунути необхідність формальної перевірки. На сьогоднішній день для такої перевірки даних запропоновано велика кількість тестів та критеріїв. Найбільш поширеними тестами можуть бути такі: тест рангової кореляції Спірмена, тест Голдфелда-Квандта, тести Глейзера іУайта.

Тест рангової кореляції Спірмена. Під час виконання цього тесту передбачається збільшення дисперсії випадкового члена, або її зменшення по мірі збільшення основної змінної x, а тому регресія, що оцінюється за допомогою методу найменших квадратів, абсолютні величини залишків моделі і значення x будуть корельованими [67]. Дані стосовно x і залишки упорядковуються, і коефіцієнт рангової кореляції та визначається так:

$$r_{x,e} = 1 - \frac{6\sum D_u^2}{n(n^2 - 1)} \tag{3.1}$$

де D – різниця між рангом x та рангом e.

Припустимо, що коефіцієнт кореляції для генеральної сукупності буде прийнято нулю, тоді коефіцієнт рангової кореляції буде мати нормальний розподіл з математичним очікуванням 0 і дисперсією 1/(n-1) на великих вибірках. Отже,

відповідна тестова статистика буде дорівнювати $r_{x,e}\sqrt{n-1}$, а при використанні двостороннього критерію нульову гіпотезу про відсутність гетероскедастичності буде відхилено на рівні значущості 5%, якщо вона перевищить 1,96, і на рівні значущості в 1 %, якщо вона перевищить 2,58. Якщо модель регресії буде включати більше однієї пояснювальної змінної, то перевірка гіпотези може виконуватися з використанням будь-якої з них. Або можна використати *t*-статистику, тоді за формулою 3.2 отримуємо:

$$t = \frac{r\sqrt{n-2}}{\sqrt{1-r^2}}, \tag{3.2}$$

де *n* – кількість спостережень; при цьому використовується поняття *df* = (*n-2*) – кількість ступенів вільності.

Знаходимо t_{kr} при кількості ступенів вільності, зазначених вище. У випадку, коли $t>t_{kr}$, гіпотеза щодо гетероскедастичності приймається. В іншому випадку, при $t<t_{kr,}$ правильним буде прийняти припущення про гомоскедастичність процесу.

Алгоритм для тесту рангової кореляції Спірмена:

1. Будуємо регресійне рівняння і оцінюємо параметри регресії.

2. Привласнюємо ранги спостереженням основної змінної *x* за зростанням.

3. Присвоюємо ранги в порядку зростання похибкам e_i (за модулем) (відхилення від лінії регресії e_i, тобто $(y_i - y)$).

4. Знаходимо D_i – різниці рангів та їх квадрати.

5. Оцінюємо значення рангового коефіцієнту кореляції.

6. Розраховуємо *t*-статистику.

7.Перевіряється гіпотеза щодо наявності або відсутності гетероскедастичності процесу.

Тест Голдфельда-Квандта. У 1956 році С. Голдфельдом і Р. Квандтом було запропоновано найбільш популярний формальний критерій для визначення гетероскедастичності [66]. При проведенні перевірки за цим критерієм передбачається, що стандартне відхилення σ_i розподілу ймовірностей u_i

пропорційне значенню змінної x у цьому спостереженні. Також передбачається, що випадковий член розподілений за нормальним розподілом без автокореляції. Усі n спостережень у вибірці необхідно упорядкувати за величиною x, після чого необхідно оцінити окремо регресії для перших n' та останніх n' спостережень; середні (n-$2n'$) відхиляються, тобто не розглядаються [66].

Якщо припущення відносно природи гетероскедастичності вірне, то дисперсія u в останніх n' спостереженнях буде більшою, ніж у перших n', і це буде виражено через суми квадратів залишків для «часткових» регресій. Позначаючи суми квадратів залишків регресії для перших n' і останніх n' спостережень відповідно через SSE_1 та SSE_2, знайдемо відношення SSR_2/SSR_1, яке має F-розподіл з (n'-k-1) і (n'-k-1) ступенями вільності, де k – кількість пояснювальних змінних у регресійному рівнянні (за їх наявності).

Потужність критерію, зазвичай, напряму залежить від вибору n' по відношенню до n. Базуючись на результатах деяких проведених Голдфельдом та Квандтом експериментів, було встановлено, що n' буде складати приблизно 11 для n=30, і близько 22, якщо n=60. Якщо в моделі є більше ніж одна пояснювальна змінна, то спостереження будуть упорядковуватися по одній з них, яка, як передбачено, пов'язана з σ_i , і n' має бути більшою, ніж k+1 (де k – кількість пояснювальних змінних) [66, 68].

Тест Глейзера. Даний тест дає можливість більш докладно розглянути характер гетероскедастичності. Ми знімаємо припущення про те, що σ_i є пропорційним x_i, і перевіряємо, чи може бути знайдена більш підходяща будь-яка інша функціональна форма, наприклад така:

$$\sigma_i = \alpha + \beta x_i^{\gamma}. \qquad (3.3)$$

Для використання даного методу необхідно оцінити регресійну залежність y від x використовуючи звичайний МНК, а потім виконати обчислення абсолютних величин залишків $|e_i|$ за функцією для даного значення γ. Кожний випадок нульової гіпотези про відсутність гетероскедастичності буде відхилено, якщо оцінка β буде значущо відрізняється від нуля. При оцінюванні більше ніж

однієї функції, отримаємо значущу оцінку β, яка свідчить про те, що орієнтиром при визначенні характеру гетероскедастичності може служити найкраща з них [69].

Тест Уайта. Цей статистичний тест дає можливість встановити чи є дисперсія залишків регресійної моделі постійною (тобто чи наявна гомоскедастичність). Перевірка дисперсії на сталість виконується шляхом регресування квадратів залишків основної регресійної моделі на добуток регресорів, квадрат регресорів та самі регресори.

У випадку, коли гомоскедастічность буде відхилена, можна використовувати моделі узагальненої авто регресії з умовною гетероскедастичністю (УАРУГ або GARCH). Цей тест і процедура оцінювання гетероскедастичності стандартних похибок, були запропоновані Халбертом Уайтом в 1980 році [67]; з тих пір вони стали широко використовуватися.

Тестування даних виконується таким чином:

1) Припускаємо, що вихідна модель має вигляд:

$$y_t = \beta_t + \beta_2 x_{2t} + \beta_3 x_{3t} + u_t. \quad (3.4)$$

У результаті оцінювання цієї моделі отримуємо регресійні залишки u_t;

2) Оцінюємо допоміжну регресію вигляду:

$$u_t = \alpha_t + \alpha_2 x_{2t} + \alpha_3 x_{3t} + \alpha_4 x^2{}_{2t} + \alpha_5 x^2{}_{3t} + \alpha_6 x_{2t} x_{3t} + \upsilon_t, \quad (3.5)$$

де υ_t – нормально розподілена похибка, незалежна від u_t.

Допоміжна регресія дає можливість визначити, чи існує якийсь систематичний зв'язок між змінами u_t і будь-якою релевантною змінною моделі (для того щоб бачити, що релевантними є саме змінні, включені в допоміжну регресію, слід представити похибку і підвести даний вираз у квадрат):

$$u_t = y_t - \beta_1 - \beta_2 x_{2t} - \beta_3 x_{3t}. \quad (3.6)$$

3) Далі досліджується статистика:

$$nR^2 \approx \chi^2 m, \quad (3.7)$$

де n – загальна кількість спостережень;

m – загальна кількість регресорів у допоміжній регресії (за винятком постійного члена) = кількості параметрів біля змінних *x*.

На основі статистики перевіримо нуль-гіпотезу H_0:

$$a_2 = 0, a_3 = 0, a_4 = 0, a_5 = 0, a_6 = 0 \quad (3.8)$$

У випадку, коли фактичні значення статистики перевищують критичні величини розподілу, нульова гіпотеза про гомоскедастичність залишків відхиляється, тобто робиться висновок про наявність гетероскедастичності [69].

3.2 Реалізація комбінованих моделей з використанням методів фільтрації

Необхідність фільтрації даних виникає кожен раз, коли потрібно відокремити передане повідомлення від шуму, який його спотворює. Мета процесу фільтрації даних, а це можуть бути не тільки результати фізичних вимірювань, а й економічні змінні діяльності, і результати соціологічних досліджень, і т. ін. – найкраще виконати відновлення початкового сигналу на тлі перешкоди (шуму), або визначати наявність корисного сигналу, або розрізнити кілька сигналів, наявних у вхідній послідовності [75].

Можна припустити, що відомі зараз методи обробки сигналів, які широко застосовуються при обробці результатів фізичних вимірювань (в радіолокації і радіотехніці), можуть знайти застосування і при створенні інформаційних систем аналізу статистичних даних та прогнозування. Пошук взаємозв’язку і закономірностей в аналізованих даних, при такому підході, повинен починатися з первинної обробки даних. Це знизить гостроту проблеми недостовірності даних, тим більше, що повністю достовірних даних просто не існує, а у будь-яких даних наявна шумова складова, яку, по можливості, слід відфільтрувати [75, 76].

3.2.1 Реалізація експоненціального згладжування

Виявлення і аналіз тенденції часового ряду часто проводиться за допомогою його вирівнювання або згладжування. Експоненціальне згладжування – один з найпростіших і розповсюджених прийомів вирівнювання ряду. Експоненціальне згладжування можна уявити як фільтр, на вхід якого послідовно надходять елементи вихідного ряду, а на виході формуються поточні значення експоненціальної середньої [76].

Припустимо, що значення у формулі (3.9) – часовий ряд.

$$X = \{x_1, ..., x_T\}. \tag{3.9}$$

Експоненціальне згладжування ряду здійснюється за рекурентною формулою:

$$S_t = \alpha x_t + (1-\alpha) S_{t-1}, \alpha \in (0,1). \tag{3.10}$$

Чим меншим є α, тим більшою мірою фільтруються, зменшуються коливання вихідного ряду і шуму. Якщо послідовно використовувати це рекурентне співвідношення, то експоненціальне середнє $S(t)$ можна виразити через значення часового ряду X:

$$S_t = \alpha x_t + (1-\alpha)(\alpha x_{t-1} + (1-\alpha) S_{t-2}) = ... = \alpha \sum_{i=0}^{t-1} (1-\alpha)^i x_{t-i} + (1-\alpha)^t S_0 \tag{3.11}$$

Якщо до моменту початку згладжування існують більш ранні дані, то в якості початкового значення S_0 можна використовувати арифметичну середню всіх наявних даних або деякої їх частини.

Після появи роботи [60] експоненціальне згладжування часто використовується для вирішення завдання короткострокового прогнозування часових рядів:

$$y_i, ..., y_t, y_i \in R. \tag{3.12}$$

Нехай задано часовий ряд

$$\hat{y}_{t+d} = f_{t,d}(y_1, ..., y_t), d \in \{1,2,...,D\}. \tag{3.13}$$

Необхідно вирішити задачу прогнозування часового ряду, тобто знайти

$$Q_T = \sum_{i=1}^{T} (y_i - \hat{y}_i) \to \min \quad (3.14)$$

Для того, щоб враховувати старіння даних, введемо незростаючу послідовність вагових коефіцієнтів:

$$w_0, w_1, ..., w_T, w_i \geq 0 \quad (3.15)$$

$$Q_T = \sum_{i=1}^{T} w_{T-i} (y_i - \hat{y}_i) \to \min \quad (3.16)$$

Якщо припустити, що D – короткостроковий прогноз, то для вирішення такого завдання можна використати модель Брауна:

$$\hat{y}_{t+d} = \alpha y_t + (1-\alpha)\hat{y}_t, \hat{y}_0 = y_0, \alpha \in (0,1) \quad (3.17)$$

Якщо розглядати прогноз на 1 крок вперед, $\hat{y}_{t+1}$, а новий прогноз виходить в результаті коригування попереднього прогнозу з урахуванням його похибки, то приходимо до адаптації процесу прогнозування [76].

При короткостроковому прогнозуванні бажано якомога швидше відбити нові зміни і в той же час якнайкраще «очистити» ряд від випадкових коливань. Таким чином слід збільшувати вагу більш останніх спостережень:

$$\alpha \to 1, \hat{y}_{t+1} \to \hat{y}_t \quad (3.18)$$

З іншого боку, для згладжування випадкових відхилень, α потрібно зменшити $\alpha \to 0, \hat{y}_{t+1} \to \hat{y}_t$. Тоді ці дві вимоги будуть знаходитись у протиріччі. Пошук компромісного значення α становить задачу оптимізації моделі. Зазвичай, α беруть з інтервалу (0,1 / 3). Модель працює тільки при невеликому горизонті прогнозування. Вона не враховує тренд і сезонні зміни. Щоб врахувати їх вплив, пропонується використовувати такі моделі: Хольта (враховується лінійний тренд) Хольта-Уінтерса (мультиплікативний експонентний тренд і сезонність), Тейла-Вейджа (адитивний лінійний тренд і сезонність) [66, 76].

3.2.2 Реалізація гранулярної фільтрації

Метод гранулярної (particle) фільтрації – це метод, в якому використовується модифікований метод Монте-Карло для розв’язання задачі

оцінювання стану досліджуваного процесу. Цей метод також відомий як алгоритм конденсації наближення взаємодіючих частинок та виживання найбільш придатних або, менш популярна назва – бутстреп фільтр. Основною ідеєю реалізації цього фільтру є представлення необхідної функції апостеріорної щільності у вигляді множини випадкових величин з відповідними ваговими коефіцієнтами та обчисленням оцінок з використанням цих величин вагових коефіцієнтів. При досягненні великої (необхідної) кількості частинок, результат застосування методу Монте-Карло досягає еквіваленту функції апріорної ймовірності, а розв'язання підходить до оптимальної оцінки Байєса [76].

Прикладна нелінійна фільтрація базується на нелінійній моделі в просторі станів у дискретному часі, яка пов'язує прихований стан x_kіз спостереженням z_k:

$$x_k = f(x_{k-1}, v_{k-1}), v_{k-1} \sim p_{v_{k-1}}, \tag{3.19}$$

$$z_k = h(x_k) + n_k, n_k \sim p_{n_k}, \tag{3.20}$$

деk– дискретний момент часу;

v_k – процес випадкового шуму з відомою щільністю розподілу p_{v_k};

n_k – адитивний шум вимірювання з відомою щільністю p_{n_k}.

Перший вимір позначається як z_1, тому перший невідомий стан це x_1, він має щільність розподілу p_{x_1}. Модель може також залежати від відомого керуючого впливу u_k, що в даному випадку опущено. Запис $s_{1:k}$ означає послідовність $s_1, s_2, \ldots\ s_k$ (тут s є одним із сигналів x, v, z, n).

У статистичній літературі модель часто записується в термінах умовної щільності розподілу таким чином:

$$x_k \sim p(x_k|x_{k-1}), \tag{3.21}$$

$$z_k \sim p(z_k|z_{k-1}). \tag{3.22}$$

Така модель є більш загальним представленням (3.19), (3.20). Апроксимація апостеріорної щільності розподілу невідомого стану за умови наявних вимірів, яка позначається як $p(x_k|z_{1:k})$, є основним завданням методів гранулярної фільтрації.

Методи гранулярної фільтрації:

1. Алгоритм послідовної вибірки за значимістю 0

Нехай $\{x_{1:k}^i, w_k^i\}_{i=1}^{N_s}$ позначає випадкову міру, яка характеризує апостеріорну щільність $p(x_{1:k}|z_{1:k})$, де $\{x_{1:k}^i, i = 0, \dots, N_s\}$ – множина k-крокових траєкторій з опорних точок (гранул, частинок) з асоційованими нормованими вагами

$\{w_k^i, i = 0, \dots, N_s\}$, $\sum_{i=1}^{N_s} w_k^i = 1$. N_s– задана кількість гранул. Тоді апостеріорна щільність в момент часу k може бути апроксимована так:

$$p(x_{1:k}|z_{1:k}) \approx \sum_{i=1}^{N_s} w_k^i \, \delta\left(x_{1:k} - x_{1:k}^i\right), \tag{3.23}$$

де $\delta(x)$ – δ-функція Дірака; тобто $p(x_{1:k}^i|z_{1:k}) \approx w_k^i$.

Вагові коефіцієнти обираються з дотриманням принципу формування вибірки за значимістю: припустимо, що $p(x)$ пропорційно$\pi(x)$ – це щільність розподілу випадкової величини, реалізацію якої з її справжнього розподілу складно згенерувати, але для якої $\pi(x)$– реальне значення, що обчислюється. Нехай також $x^i \sim q(x), i = 0, \dots, N_s$– реалізації випадкової величини, які легко генеруються з розподілу зі щільністю $q(\cdot)$, яка називається пропонуючою щільністю або значимою щільністю (proposal density, importance density). Тоді зважена апроксимація щільності $p(\cdot)$ виглядає таким чином:

$$p(x) \approx \sum_{i=1}^{N_s} w^i \, \delta\left(x - x^i\right), \tag{3.24}$$

де

$$w^i \propto \frac{\pi\left(x^i\right)}{q(x^i)}. \tag{3.25}$$

є нормованою вагою i-ої гранули (після обчислення відношення виконується нормування ваг: $\sum_{i=1}^{N_s} w_k^i = 1$).

Тоді, якщо реалізації $x_{1:k}^i$ були згенеровані з розподілу з пропонуючою щільністю $q(x_{1:k}|z_{1:k})$, то ваги в (3.24) відповідно до (3.25) такі:

$$w^i \propto \frac{p(x_{1:k}|z_{1:k})}{q(x_{1:k}|z_{1:k})}. \tag{3.26}$$

У випадку послідовних обчислень, на кожній ітерації маючи зважену вибірку, яка апроксимує $p(x_{1:k-1}|z_{1:k-1})$, можна було б апроксимувати $p(x_{1:k}|z_{1:k})$ новою вибіркою. Але якщо пропонуюча щільність обрана так, що вона розкладалась на множники:

$$q(x_{1:k}|z_{1:k}) = q(x_k|x_{1:k-1}, z_{1:k})q(x_{1:k-1}|z_{1:k-1}), \tag{3.27}$$

то можна отримати елементи $x_{1:k}^i \sim q(x_{1:k}|z_{1:k})$,доповнюючи кожен з існуючих елементів $x_{1:k-1}^i \sim q(x_{1:k-1}|z_{1:k-1})$ новим станом $x_k^i \sim q(x_k|x_{1:k-1}, z_{1:k})$.

У практичних застосуваннях найчастіше необхідна лише фільтрована оцінка розподілу $p(x_k|z_{1:k})$ на кожному кроці, а не умовний розподіл одразу всієї траєкторії $p(x_{1:k}|z_{1:k})$. Тому надалі будемо розглядати саме цей випадок.

Використовуючи прямий наслідок з теореми Байєса

$$p(x_k|z_{1:k}) = \frac{p(z_k|x_k)p(x_k|z_{1:k-1})}{p(z_k|z_{1:k-1})}, \tag{3.28}$$

за виконання певних умов, зокрема, якщо $q(x_k|x_{1:k-1}, z_{1:k})$=$q(x_k|x_{k-1}, z_k)$, тобто коли пропонуюча щільність залежить лише від x_{k-1} та z_k, можна показати [15], що для оновлення ваг використовується таке співвідношення:

$$w_k^i \propto w_{k-1}^i \frac{p\left(z_k\middle|x_k^i\right)p\left(x_k^i\middle|x_{k-1}^i\right)}{q\left(x_k^i\middle|x_{k-1}^i, z_k\right)}, \tag{3.29}$$

і фільтрований апостеріорний розподіл може бути апроксимований так:

$$p(x_k|z_{1:k}) \approx \sum_{i=1}^{N_s} w_k^i \, \delta\left(x_k - x_k^i\right). \tag{3.30}$$

Вибір пропонуючого розподілу є одним із найбільш важливих етапів

проектування гранулярного фільтра. Зокрема важливо, щоб дисперсія вагових коефіцієнтівw_k^i була невеликою. Найчастіше в якості пропонуючого розподілу використовують апріорний такого типу (3.30):

$$q\left(x_k\middle|x_{k-1}^i, z_k\right) = p\left(x_k\middle|x_{k-1}^i\right), \tag{3.31}$$

що є зручним. В такому разі (3.31) спрощується до вигляду:

$$w_k^i \propto w_{k-1}^i p\left(z_k\middle|x_k^i\right). \tag{3.32}$$

Однак, не в усіх задачах такий вибір пропонуючого розподілу є оптимальним. Підсумуємо алгоритм послідовної вибірки за значимістю.

Елементи $x_{1:1}^i$у зваженій вибірці на 1-му кроці $\left\{x_{1:1}^i, \frac{1}{N_s}\right\}_{i=1}^{N_s}$ генеруються з початкового розподілу p_{x_1}. Оскільки цей розподіл приймається як істинний, то робити корекції не потрібно і всі вагові коефіцієнти рівні $w_1^i = \frac{1}{N_s}$.

Процедуру генерування зваженої вибірки на k-му кроці, маючи зважену вибірку на $(k-1)$-му кроці, можна представити таким псевдокодом:

Algorithm 1: SIS Particle Filter

$\left[\left\{x_k^i, w_k^i\right\}_{i=1}^{N_s}\right]$= SIS$\left[\left\{x_{k-1}^i, w_{k-1}^i\right\}_{i=1}^{N_s}, z_k\right]$

FOR

— Згенерувати $x_k^i \sim q\left(x_k\middle|x_{k-1}^i, z_k\right)$;

— Присвоїти гранулі x_k^iвагуw_k^i відповідно до (3.32).

ENDFOR

Для цього та всіх подальших алгоритмів фільтрації на кожному кроці апостеріорний розподіл можна апроксимувати за формулою (3.32). Оцінкою умовного математичного очікування стану x_kбуде таке значення:

$$\widehat{x_k} = \sum_{i=1}^{N_s} w_k^i x_k^i \,. \tag{3.33}$$

1. Відсіювання. Однією з проблем гранулярної фільтрації є виродження

вагів, тобто коли після певної кількості ітерацій усі гранули, крім однієї, будуть мати незначну вагу. Дисперсія вагів може лише зростати з часом, тому феномену виродження уникнути неможливо. Виродження призводить до того, що багато обчислень витрачається на оновлення гранул, внесок яких у апроксимацію $p(x_k|z_{1:k})$ майже нульовий.

Одним із підходів для запобігання ефекту виродження є проведення відсіювання (resampling). Основна ідея цього методу відсіювання є виключення гранул, що мають малу вагу та фокусування на гранулах з великою вагою. На кроці відсіювання відбувається генерування нової множини $\{x_k^{i*}\}_{i=1}^{N_s}$ вибором (з повтореннями) N_s реалізацій випадкової величини x_k з наближеного дискретного представлення розподілу $p(x_k|z_{1:k})$, заданого формулою (3.32), тобто таким чином, що $P\{x_k^{i*} = x_k^j\} = w_k^j$. Отримана вибірка є насправді вибіркою незалежних однаково розподілених величин з дискретного розподілу (3.33), тому вагові коефіцієнти тепер встановлюються рівними, $w_k^i = \frac{1}{N_s}$.

Наведемо алгоритм процедури відсіювання. Такий варіант відсіювання, систематичний відсів (systematic resampling) було обрано через простоту реалізації та оцінку складності алгоритму $O(N_s)$. Для кожної частинки нової вибірки зберігається також її індекс у попередній вибірці, i^j, який знадобиться в наступних алгоритмах.

Algorithm 2: Resampling Algorithm

$\left[\{x_k^{j*}, w_k^j, i^j\}_{j=1}^{N_s}\right]$= RESAMPLE$\left[\{x_k^i, w_k^i\}_{i=1}^{N_s}\right]$

Ініціалізувати функцію розподілу (ФР): $c_1 = 0$;

FOR$i = \overline{2, N_s}$

— Побудувати ФР:$c_i = c_{i-1} + w_k^i$;

ENDFOR

Почати з початку ФР: $i = 1$

Згенерувати початкову точку:$u_1 \sim U[0, N_s^{-1}]$

FOR$j = \overline{1, N_s}$

— Рухатись уздовж ФР: $u_j = u_1 + N_s^{-1}(j-1)$;

— WHILE$u_j > c_i$

— — $i = i + 1$

— ENDWHILE

— Присвоїти елемент: $x_k^{j*} = x_k^i$;

— Присвоїти вагу: $w_k^j = N_s^{-1}$;

— Присвоїти: $i^j = i$

ENDFOR

Окрім переваг, процедура відсіювання також має недоліки. По-перше, звуження можливості для паралелізму, крок відсіювання передбачає комбінування всіх гранул. Через те, що деякі гранули можуть мати велику вагу та обиратися багаторазово, втрачається різноманітність вибірки, а у результуючій вибірці може бути дуже багато повторів. Таку проблему ще називають збідненням вибірки у випадку, коли є малий шум процесу. Це може призвести до того, що всі гранули будуть поєднані в одну через декілька пройдених ітерацій.

2. Фільтр вибірки за значимістю з відсівом. Це метод Монте-Карло, що може застосовуватися до розв’язання задач рекурсивної байєсівської фільтрації. Обмеження, накладені на його використання, є дуже слабкими. Функції $f(\cdot,\cdot)$ та $h(\cdot,\cdot)$ в (3.32) і (3.33) мають бути відомими, але необхідно мати змогу генерувати реалізації з розподілу шумової складової$p_{v_{k-1}}$ та апріорного розподілу $p(x_k|x_{k-1})$ і визначати значення щільності розподілу $p(z_k|x_k)$ в потрібних точках, принаймні з точністю до спільної константи. SIR-алгоритм може бути легко виведений з алгоритму SIS за рахунок відповідного вибору [76]:

– пропонуючого розподілу $q(x_k|x_{k-1}^i, z_k)$ – в якості нього обирається апріорна щільність $p\left(x_k\middle|x_{k-1}^i\right)$;

– кроку відсіювання, який здійснюється у кожний момент часу.

Такий вибір пропонуючого розподілу обумовлює необхідність

генерування реалізацій з $p(x_k|x_{k-1}^i)$. Реалізацію $x_k^i \sim p(x_k|x_{k-1}^i)$ можна отримати, якщо згенерувати реалізацію шумової складової,$v_{k-1}^i \sim p_{v_{k-1}}$, та покласти $x_k^i = f(x_{k-1}^i, v_{k-1}^i)$.

Для такого спеціального вибору пропонуючої щільності формула оновлення вагових коефіцієнтів набуває вигляду (3.32). Але, взявши до уваги, що відсів відбувається в кожний момент часу, отримуємо, що $w_{k-1}^i = \frac{1}{N_s} \ \forall i$, а тому:

$$w_k^i \propto p(z_k|x_k^i). \tag{3.34}$$

Ваги нормуються перед фазою відсіювання. Наведемо псевдокод ітерації цього алгоритму:

Algorithm 3: SIR Particle Filter

$\left[\{x_k^i, w_k^i\}_{i=1}^{N_s}\right]$ = SIR$\left[\{x_{k-1}^i, w_{k-1}^i\}_{i=1}^{N_s}, z_k\right]$

FOR$i = \overline{1, N_s}$

— Згенерувати $x_k^i \sim p(x_k|x_{k-1}^i)$

— Обчислити$w_k^i = p(z_k|x_k^i)$;

ENDFOR

Обчислити загальну вагу: $t = \sum_{i=1}^{N_s} w_k^i$;

FOR$i = \overline{1, N_s}$

— Нормувати і-ту вагу:$w_k^i = t^{-1} w_k^i$;

ENDFOR

Провести відсіювання, використовуючи алгоритм 2 (Resampling Algorithm):

— $\left[\{x_k^i, w_k^i, -\}_{i=1}^{N_s}\right]$ = RESAMPLE$\left[\{x_k^i, w_k^i\}_{i=1}^{N_s}\right]$.

3. Додатковий фільтр вибірки за значимістю з відсівом

Додатковий фільтр вибірки за значимістю з відсіюванням (Auxiliary Sampling Importance Resampling Filter (ASIR)) є одним з варіантів стандартного фільтра вибірки за значимістю з відсіюванням. Його реалізація може бути

отримана з SIR фільтра введенням пропонуючої щільності,$q(x_k, i|z_{1:k})$, з якої генеруються реалізації пар,$\{x_k^j, i^j\}_{j=1}^{N_s}$, де i^j позначає номер гранули в $(k-1)$-й момент.i^j в даному фільтрі називають допоміжною змінною, оскільки її єдина мета допомогти в задачі симуляції; звідси походить назва фільтра.

Використовуючи теорему Байєса, можна показати, що

$$p(x_k, i|z_{1:k}) \propto p(z_k|x_k)p\left(x_k\middle|x_{k-1}^i\right)w_{k-1}^i. \tag{3.35}$$

ASIR фільтр генерує реалізацію зі спільного розподілу $p(x_k, i|z_{1:k})$, а потім опускає номер i в парі (x_k, i), щоб отримати вибірку $\{x_k^j\}_{j=1}^{N_s}$ з маргінального розподілу $p(x_k|z_{1:k})$. Пропонуюча щільність для $\{x_k^j, i^j\}_{j=1}^{N_s}$ має задовольняти пропорційність:

$$q(x_k, i|z_{1:k}) \propto p\left(z_k\middle|\mu_k^i\right)p\left(x_k\middle|x_{k-1}^i\right)w_{k-1}^i, \tag{3.36}$$

де μ_k^i – певна характеристика x_k за умови x_{k-1}^i. Наприклад, це може бути умовне математичне очікування$\mu_k^i = E\left[x_k|x_{k-1}^i\right]$, модаабо реалізація $\mu_k^i \sim p(x_k|x_{k-1}^i)$.

Якщо записати (3.35) та покласти (3.36):

$$q(x_k, i|z_{1:k}) = q(x_k|i, z_{1:k})q(i|z_{1:k}) \tag{3.37}$$

$$q(x_k|i, z_{1:k}) = p(x_k|x_{k-1}^i) \tag{3.38}$$

то з формул (3.37) та (3.38) випливає формула (3.39):

$$q(i|z_{1:k}) \propto p\left(z_k\middle|\mu_k^i\right)w_{k-1}^i \tag{3.39}$$

Парі $\{x_k^j, i^j\}_{j=1}^{N_s}$ присвоюється вага, пропорційна відношенню

$$w_k^j \propto \frac{p\left(z_k\middle|x_k^j\right)}{p\left(z_k\middle|\mu_k^{i^j}\right)}. \tag{3.40}$$

Наведемо псевдокод ітерації цього алгоритму:

Algorithm 4: Auxiliary Particle Filter

$\left[\{x_k^i, w_k^i\}_{i=1}^{N_s}\right]$ = APF$\left[\{x_{k-1}^i, w_{k-1}^i\}_{i=1}^{N_s}, z_k\right]$

FOR$i = \overline{1, N_s}$

— Обчислити μ_k^i

— Обчислити$w_k^i = q(i|z_{1:k}) \propto p(z_k|\mu_k^i)w_{k-1}^i$.

ENDFOR

Обчислити загальну вагу: $t = \sum_{i=1}^{N_s} w_k^i$;

FOR$i = \overline{1, N_s}$

— Нормувати і-ту вагу:$w_k^i = t^{-1}w_k^i$;

ENDFOR

Провести відсів, використовуючи алгоритм 2 (ResamplingAlgorithm):

— $\left[\{—, —, i^j\}_{j=1}^{N_s}\right]$ = RESAMPLE$\left[\{x_{k-1}^i, w_k^i\}_{i=1}^{N_s}\right]$

FOR$j = \overline{1, N_s}$

— Згенерувати$x_k^j \sim q\left(x_k \middle| i^j, z_{1:k}\right) = p(x_k|x_{k-1}^{i^j})$, як в SIR-фільтрі

— Присвоїти вагу w_k^j;

ENDFOR

Обчислити загальну вагу: $t = \sum_{i=1}^{N_s} w_k^i$;

FOR$i = \overline{1, N_s}$

— Нормувати і-ту вагу:$w_k^i = t^{-1}w_k^i$;

ENDFOR

Таким чином, ідейно, щоб згенерувати реалізацію з розподілу $p(x_k, i|z_{1:k})$, спочатку генерується номер i з ймовірністю $w_k^i \propto q(i|z_{1:k})$ (ці ймовірності називають ймовірностями першого етапу), а потім x_k з розподілу $p(x_k|x_{k-1}^i)$. Після цього індекс i відкидається, а x_k присвоюються відповідні ваги. При такому підході ваги другого рівня можуть мати меншу дисперсію ніж оригінальні ваги, отримані за SIR-алгоритмом.

4. Регуляризація гранулярного фільтру

Метод відсіювання було введено для зменшення ефекту виродження гранулярного фільтру. Як було відмічено, відсів може призвести до інших проблем, зокрема проблеми втрати різноманітності гранул. Ця проблема виникає через те, що на етапі відсіювання реалізації генерація виконується з дискретного, а не неперервного розподілу. Якщо цю проблему залишити без уваги та не відслідковувати її, вона може призвести до екстремального випадку «зліплювання гранул», коли всі N_sгранул будуть знаходитися в одному й тому самому місці простору станів і погано відображати апостеріорний розподіл [72, 77].

Для передбачення цієї проблеми було запропоновано регуляризований гранулярний фільтр (Regularized Particle Filter (RPF)). Під час проведення процедури відсіювання генеруються гранули з неперервної апроксимації $p(x_k|z_{1:k})$, а не з дискретної, як в SIR-фільтрі. А саме апроксимації вигляду:

$$p(x_k|z_{1:k}) \approx \sum_{i=1}^{N_s} w_k^i \, K_h\left(x_k - x_k^i\right), \tag{3.41}$$

де

$$K_h(x) = \frac{1}{h^{n_x}} K\left(\frac{x}{h}\right) \tag{3.42}$$

це масштабоване статистичне ядро (kernel density) $K(\cdot)$;

$h > 0$ – це параметр згладжування (скалярний);

n_x – розмірність вектора стану x;

$w_k^i, i = 1, \dots, N^s$ – нормовані ваги.

Статистичне ядро $K(\cdot)$– це симетрична функція щільності розподілу така, що:

$$\int xK(x)dx = 0, \int \|x\|^2 K(x)dx < \infty \, . \tag{3.43}$$

Тобто випадковий вектор з щільністю $K(\cdot)$ має нульове математичне очікування та скінченні моменти другого порядку.

Статистичне ядро $K(\cdot)$ та параметр згладжування $h > 0$необхідно обрати такими, щоб середнє інтегроване квадратичне відхилення було мінімізованим

(mean integrated square error (MISE)) на значеннях справжньої апостеріорної щільності та відповідним регуляризованим емпіричним представленням в (3.43), яке визначається так:

$$MISE(\hat{p}) = E[\int(\hat{p}(x_k|z_{1:k}) - p(x_k|z_{1:k}))^2 dx_k] , \quad (3.44)$$

де$\hat{p}(x_k|z_{1:k})$позначає апроксимацію, визначену в (3.44). У спеціальному випадку, коли всі елементи мають однакову вагу, $\frac{1}{N_s}$, оптимальним вибором ядра є ядро Єпанечнікова:

$$K_e(x) = \begin{cases} \frac{n_x + 2}{2c_{n_x}}(1 - \|x\|^2), & \text{якщо } \|x\| < 1 \\ 0, & \text{інакше} \end{cases}, \quad (3.45)$$

де c_{n_x} – об’єм одиничної гіперсфери в просторі R^{n_x}.

Часто зручно використовувати гаусівське ядро:

$$K_n(x) = (2\pi)^{-\frac{n_x}{2}} e^{-\frac{1}{2}\|x\|^2}. \quad (3.46)$$

Крім того, у випадку, коли щільність, що апроксимується, є гаусівською з одиничною коваріаційною матрицею, тоді оптимальним вибором параметра згладжування є такий:

$$h_{opt} = A(K)N_s^{1/(n_x+4)}, \quad (3.47)$$

де $A(K_e) = \left[8c_{n_x}^{-1}(n_x + 4)(2\sqrt{\pi})^{n_x}\right]^{1/(n_x+4)}$для ядра Єпанечнікова та

$A(K_n) = [4/(n_x + 2)]^{1/(n_x+4)}$для гаусівського ядра.

В реалізації алгоритму регуляризованого гранулярного фільтра відсіювання буде проводиться не на кожному кроці, а якщо значення міри виродження стає нижчим деякого заданого порогу N_T. Сама міра виродження обчислюється за формулою:

$$N_{eff} = \frac{1}{\sum_{i=1}^{N_s}(w_k^i)^2} . \quad (3.48)$$

Зазначимо, що якщо встановити порогове значення N_T рівнимN_s, то відсів відбуватиметься на кожній ітерації.

Наведемо псевдокод ітерації алгоритму:

Algorithm 5: Regularized Particle Filter

$\left[\left\{x_k^{i*}, w_k^i\right\}_{i=1}^{N_s}\right]$ = RPF$\left[\left\{x_{k-1}^i, w_{k-1}^i\right\}_{i=1}^{N_s}, z_k\right]$

FOR$i = \overline{1, N_s}$

— Згенерувати $x_k^i \sim q(x_k | x_{k-1}^i, z_k)$;

— Присвоїти гранулі x_k^iвагуw_k^i;

ENDFOR

Обчислити загальну вагу: $t = \sum_{i=1}^{N_s} w_k^i$;

FOR$i = \overline{1, N_s}$

— Нормувати і-ту вагу:$w_k^i = t^{-1} w_k^i$;

ENDFOR

ОбчислитиN_{eff},

IF$N_{eff} < N_T$

— Обчислити емпіричну коваріаційну матрицю S_k для $\left\{x_k^i, w_k^i\right\}_{i=1}^{N_s}$

— Обчислити матрицю Σ_kтаку, що $\Sigma_k \Sigma_k^T = S_k$ (наприклад, $\Sigma_k = E_k D_k^{\frac{1}{2}}$, де $S_k = E_k D_k \mathrm{E}_k^T$ – спектральний розклад матриці S_k)

— Провести відсів, використовуючи алгоритм 2 (Resampling Algorithm):

— — $\left[\left\{x_k^i, w_k^i, -\right\}_{i=1}^{N_s}\right]$ = RESAMPLE$\left[\left\{x_k^i, w_k^i\right\}_{i=1}^{N_s}\right]$

— FOR$i = \overline{1, N_s}$

— — Згенерувати$\varepsilon^i \sim K$зі статистичного ядра

— — $x_k^{i*} = x_k^i + h_{opt} \Sigma_k \varepsilon^i$;

— ENDFOR

ENDIF

Результати будуть оптимальними лише у випадках, коли вони можуть використовуватись в загальному представленні для отримання субоптимального фільтра. Введення регуляризації дає покращення результату у порівнянні зі звичайним SIR-фільтром у тих випадках, коли наявна значна втрата різноманітності вибірки гранул, наприклад, якщо шум процесу є незначним.

Проблемою такого підходу буде збільшення дисперсії апостеріорного розподілу.

3.3 Моделі гетероскедастичних процесів

Важливим моментом при моделюванні динаміки системи є виявлення та визначення типу можливих невизначеностей. Невизначеності будемо розглядати як фактори негативного впливу на процес моделювання, що можуть призвести до різних помилок, зокрема зниження якості кінцевих результатів. Лінійні процеси можуть бути нестаціонарними, якщо вони містять лінійний тренд. Зазвичай, нелінійні процеси також можуть бути частково стаціонарними, але в основному мати стабільний режим роботи [70]. Нелінійні нестаціонарні процеси (ННП) часто представлені у різноманітних областях досліджень. До них можуть відноситися нелінійний інтегрований процес з трендами другого і більше порядку, коінтегровані процеси з однаковим ступенем інтеграції і гетероскедастичні процеси. Гетероскедастичні процеси передбачають одночасну побудову двох типів моделей: моделі самого процесу (амплітуди) та моделі опису динаміки умовної дисперсії, яка широко використовується на практиці для розв'язання задач діагностики (технічної, медичної, фінансово-економічної), аналізу ризиків у різних сферах, біржової торгівлі та ін. Адаптивне прогнозування нелінійних нестаціонарних процесів на сьогодні є дуже важливою задачею сучасності, у зв'язку з тим, що велика кількість процесів в економіці, фінансах, екології та технологічних процесах можна віднести до вказаного класу. Розглянемо деякі відомі математичні моделі нелінійних нестаціонарних процесів [72].

Поліноміальна регресія. Одним з часто використовуваних методів опису з трендом є поліноміальна регресія такого типу:

$$y(k) = a_0 + a_1 k + a_2 k^2 + \ldots + a_m k^m + \varepsilon(k), \tag{3.49}$$

де $y(k)$ – основна (залежна) змінна процесу;

$k = 0, 1, \ldots$ – дискретний час, пов'язаний з реальним неперервним часом t через період дискретизації значень змінних;

$a_i, i = 0$ – коефіцієнти (параметри) моделі;

m – порядок поліному, визначений кількістю похідних, які можна розрахувати на основі адекватної поліноміальної моделі процесу;

$\varepsilon(k)$ – випадковий процес, який зумовлений наявністю випадкових зовнішніх збурень, похибок вимірів, методичних похибок оцінок параметрів моделі і помилок оцінювання її структури.

Поліноми довільного порядку часто використовують на практиці через простоту визначення структури моделі і можливості використовувати методу найменших квадратів (МНК) для оцінки параметрів.

Авторегресія з трендовою складовою:

$$y(k) = a_0 + \sum_{i=1}^{p} a_i y(k-i) + \sum_{j=1}^{m} b_i k^j + \varepsilon(k) \quad (3.50)$$

.

Авторегресія з інтегрованим ковзним середнім:

$$y(k) = a_0 + \sum_{i=1}^{p} a_i y(k-i) + \sum_{j=1}^{q} b_j v(k-j) + \varepsilon(k) \quad (3.51)$$

.

За формулами (3.50) та (3.51) описано тренд і коливання, які на нього накладаються. Оскільки процеси такого роду є досить характерними для виробничих технологій, економіки, фінансів, екології та інших галузей, то їм приділяється значна увага. Нестаціонарні процеси такого типу дуже часто можна зустріти в економіці і відповідній фінансовій діяльності, які характеризуються високою нестаціонарною динамікою розвитку.

Авторегресійна умовна гетероскедастичність АРУГ (Auto Regressive Conditional Heteroscedasticity). Дана модель використовується в економетриці при аналізі часових рядів (в першу чергу фінансових), у яких умовна дисперсія ряду напряму залежить від минулих значень, значень дисперсій та інших факторів. АРУГ моделі призначені для кластеризації волатильності фінансових ринків, коли періоди високої волатильності можуть займати деякий час, змінюючись потім періодами низької волатильності, причому середню (довгострокову, безумовну) волатильність можна вважати відносно стабільною [72].

Авторегресійне умовно гетероскедастичне(АРУГ) рівняння має вигляд:

$$\hat{\varepsilon}^2(k) = \alpha_0 + \alpha_1\hat{\varepsilon}^2(k-1) + \alpha_2\hat{\varepsilon}^2(k-2) + \cdots + \alpha_q\hat{\varepsilon}^2(k-q) + v(k), \quad (3.52)$$

де $\hat{\varepsilon}^2(k)$ – квадрати оцінок залишків (похибок) моделі;

α_0 – коефіцієнт затримки;

$\alpha_1, \dots, \alpha_q$ – параметри;

$v(k)$ – процес білого шуму з нульовим середнім для адекватної моделі.

Залишки (збурення) $\varepsilon(k)$ можуть бути отримані на основі рівнянь регресії, авторегресії або авторегресії з ковзним середнім низького порядку.

Додатково до рівняння (3.52) можна вибрати і більш складніші форми описання поведінки дисперсії. Наприклад, наперед майже ніколи невідомо як може вплинути збурення на процес – мультиплікативно чи адитивно. Тому його можна представити у моделі в мультиплікативній формі:

$$\varepsilon^2(k) = v^2(k)[\alpha_0 + \alpha_1\varepsilon^2(k-1)], \quad (3.53)$$

де$v(k)$ – мультиплікативне збурення у формі білого шуму, причому $\{v(k)\}\sim(0,1)$ має нульове середнє і одиничну дисперсію;

$\varepsilon(k-1)$ і $v(k)$ – статистично незалежні (некорельовані) величини.

Основним недоліком АРУГ є те, що $\alpha_i, i = 0, \dots q$ мають бути невід'ємними, щоб умовна дисперсія завжди була позитивною.

Узагальнена авторегресійна умовно гетероскедастична модель УАРУГ(Generalized Autoregressive ConditionalHeteroscedastic) передбачає, що для поточної зміни дисперсії будуть впливати попередні оцінки дисперсії та зміни показників [71]. Розширення АРУГ моделі є описання умовної дисперсії як процесу АРКС. Нехай похибки описуються рівнянням

$$\varepsilon(k) = v(k)\sqrt{h(k)}, \quad (3.54)$$

де$\sigma_v^2 = 1$;

$h(k)$ – умовна дисперсія, яка визначається за виразом:

$$h(k) = \alpha_0 + \sum_{i=1}^{q} \alpha_i \varepsilon^2(k-i) + \sum_{i=1}^{p} \beta_i h(k-i), \tag{3.55}$$

де p – кількість попередніх оцінок, які впливають на поточне значення;

β_i – вагові коефіцієнти, які відображають ступінь впливу попередніх оцінок на поточне значення.

УАРУГ(p, q) має в складі компоненти авторегресії та ковзного середнього відносно дисперсії гетероскедастичного процесу. По відношенню до узагальнених умовно гетероскедастичних процесів не узгоджується визначення стаціонарності, тобто може бути так, що сильно стаціонарний процес УАРУГ не завжди буде слабо стаціонарним, або навпаки. Через це виникає проблема знаходження та визначення стаціонарності в таких процесах[71, 72].

Експоненційна узагальнена авторегресійна умовно гетероскедастична модель ЕУАРУГ (Exponential Generalized Autoregressive Conditional Heteroscedastic або EGARCH) не включає в собі недоліки АРУГ та УАРУГ моделей. В даній моделі логарифм умовної дисперсії може визначатися з використанням функції нормованих похибок g(·):

$$\log[h(k)] = c_0 + \sum_{i=1}^{\infty} c_i \mathrm{g}[y(k-i)] \tag{3.56}$$

$$\mathrm{g}(y) = \alpha y(k) + \beta[|y(k)| - E|y(k)|], \tag{3.57}$$

де $E[\mathrm{g}(y)] = 0$;

α, β – параметри моделі;

$y(k)$ – основна змінна, що моделюється.

– Узагальнена білінійна модель. Наведена модель широко застосовується і має дуже зручну структуру:

$$y(k) = a_0 + \sum_{i=1}^{p} a_i y(k-i) + \sum_{j=1}^{q} b_j v(k-j) + + \sum_{i=1}^{m} \sum_{j=1}^{s} c_{i,j} y(k-i) v(k-j) + \varepsilon(k), \tag{3.58}$$

де p, q, m і sє додатними числами, які відображають порядок моделі.

– Лінійна комбінація лінійних та нелінійних компонентів.

Найчастіше моделювання нелінійних процесів ґрунтується на лінійній комбінації лінійних та нелінійних компонентів:

$$y(k) = \beta^T z(k) + \sum_{i=1}^{p} \alpha_i \varphi_i(\theta_i^T z(k)) + \varepsilon(k), \quad (3.59)$$

де $z(k)$ – вектор значень затримки часу залежної змінної $y(k)$, а також попередні та поточні значення незалежних пояснювальних змінних$\boldsymbol{x}(k)$ з відповідним зміщенням часу[72];

φ_i– набір (лінійних та нелінійних) функцій, які включають наступні компоненти: функцію потужності$\varphi_i(x) = x^i$, тригонометричні функції $\varphi_i(x) = \sin x$або$\varphi_i(x) = \cos x$та ін.

– Логіт і пробіт моделі

Припустимо, що

$$P\{y_i = 1\} = F(\beta^T x_i), \quad (3.60)$$

де$F(x)$ – деяка функція, область значень якої лежить у відрізку $[0,1]$;

y_i – бінарна змінна, тобто може приймати значення 0 або 1.

Якщо у якості функції $F(x)$ використовують такі:

а) функцію стандартного нормального розподілу

$$\Phi(u) = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{u} e^{-\frac{x^2}{2}} dz, \quad (3.61)$$

тоді модель називають пробіт моделлю;

б) функцію логістичного розподілу

$$\Lambda(u) = \frac{e^u}{1 + e^u}, \quad (3.62)$$

тоді модель називають логіт моделлю.

Для оцінювання параметрів β у логіт та пробіт моделях зазвичай використовують метод максимальної правдоподібності, припускаючи, що спостереження $y_1, \dots, y_n$ незалежні. Оскільки y_i може приймати значення 0 або 1, то функція правдоподібності має такий вигляд:

$$\mathcal{L}(y_1, \dots, y_n) = \prod_{i:y_i=0} \left(1 - F(\beta^T x_i)\right) \prod_{i:y_i=1} F(\beta^T x_i) == \prod_{i=1}^{n} F^{y_i}(\beta^T x_i)\left(1 - F(\beta^T x_i)\right)^{1-y_i} \tag{3.63}$$

Диференціюванням логарифмічної функції правдоподібності по вектору β отримаємо рівняння правдоподібності у векторній формі [71, 72]:

$$\frac{\partial \ln \mathcal{L}}{\partial \beta} = \sum_{i=1}^{n} \left(\frac{y_i f(\beta^T x_i)}{F(\beta^T x_i)} - \frac{(1-y) f(\beta^T x_i))}{1 - F(\beta^T x_i)} \right) x_i = 0, \tag{3.64}$$

де$f(x)$ – щільність розподілу, яка відповідає функції $F(x)$.

Використовуючи те, що $\Lambda'(u) = \Lambda(u)(1 - \Lambda(u))$,отримаємо таке рівняння правдоподібності для логіт моделі:

$$\sum_{i=1}^{n} \left(y_i - \Lambda(\beta^T x_i)\right) x_i = 0 \,. \tag{3.65}$$

Для пробіт моделі логарифмічну функцію правдоподібності можна записати у наступному вигляді:

$$\ln \mathcal{L} = \sum_{i:y_i=0} \ln(1 - \Phi(\beta^T x_i)) + \sum_{i:y_i=1} \ln \Phi(\beta^T x_i) \tag{3.66}$$

Тоді рівняння правдоподібності буде таким:

$$\frac{\partial \ln \mathcal{L}}{\partial \beta} = \sum_{i:y_i=0} \frac{-\varphi(\beta^T x_i)}{1 - \Phi(\beta^T x_i)} x_i + \sum_{i:y_i=1} \frac{\varphi(\beta^T x_i)}{\Phi(\beta^T x_i)} x_i, \tag{3.67}$$

де$\varphi(x) = \Phi'(x)$.

Оскільки нормальний розподіл, як і логістичний, симетричний, $1 - \Phi(\beta^T x_i) == \Phi(-\beta^T x_i)$, то отримаємо:

$$\frac{\partial \ln \mathcal{L}}{\partial \beta} = \sum_{i=1}^{n} \frac{q_i \varphi(\beta^T x_i)}{\Phi(q_i \beta^T x_i)} x_i = \sum_{i=1}^{n} \lambda_i \, x_i = 0, \tag{3.68}$$

де

$$q_i = 2y_i - 1, \qquad \lambda_i = \frac{q_i \varphi(\beta^T x_i)}{\Phi(q_i \beta^T x_i)} \tag{3.69}$$

– Лінійна комбінація випадкових процесів (випадкові тренди)

Випадкові тренди можуть бути описані комбінаціями випадкових величин

з параметрами, що визначаються за допомогою характеристик фактичних випадкових процесів. В загальному випадку процес може складатися як з детермінованої, так і випадкової складової.

Модель випадкового кроку є однією з найпростіших, яка дозволяє описати випадковий тренд в деяких випадках. Вона має вигляд:

$$y(k) = y(k-1) + \varepsilon(k), \tag{3.70}$$

або

$$\Delta y(k) = \varepsilon(k), \tag{3.71}$$

де $\varepsilon(k)$ – є білим шумом із нульовим середнім.

3.4 Параметри моделей гетероскедастичних процесів

Процес оцінювання параметрів моделі представляє собою задачу складність якої зростає зі складністю моделі. Короткий огляд різних методів оцінювання параметрів моделей наведено нижче. Вершину ієрархії представляє оцінювання за методом максимальної правдоподібності; оцінювання ґрунтується на мінімізації логарифма функції правдоподібності:

$$-\ln l(\theta) = -\ln p(\xi^0, \xi^1, ..., \xi^n \mid \theta), \tag{3.72}$$

де θ – вектор невідомих параметрів; $(\{\xi\} \mid \theta)$ і p – щільність імовірності спостережень процесу і ξ при фіксованому параметрі θ . Якщо процес марковський, то формула спрощується до вигляду:

$$-\ln l(\theta) = -\ln p(\xi^0 \mid \theta) - \sum_{i=0}^{n-1} p(\xi^{i+1}, \theta), \tag{3.73}$$

де $p(\xi^{i+1} \mid \xi^i, \theta)$ – щільність перехідної ймовірності за умови відомого стану і ξ і параметра θ. Аналітичний вираз для цієї ймовірності отримується досить рідко, та призводить до розробки широкого спектру методів, які не вимагають аналітичного представлення для перехідної щільності.

Відомо, що щільність перехідної ймовірності задовольняє рівнянню Фоккера-Планка [54].

$$\frac{\partial f}{\partial t} = \left[-\sum_{i=1}^{N} \frac{\partial}{\partial x_i} D_i^1(x_1,...,x_N) + \sum_{i=1}^{N}\sum_{j=1}^{N} \frac{\partial^2}{\partial x_i \partial x_j} D_{ij}^2(x_1,...,x_N) \right] f, \quad (3.74)$$

$$D_i^1(x,t) = \mu_i(x,t), \quad (3.75)$$

$$D_{ij}^2(x,t) = \frac{1}{2}\sum_k \sigma_{ik}(x,t)\sigma_{jk}(x,t) \quad (3.76)$$

.

з відповідними початковими і граничними умовами. У найпростіших випадках воно повинно вирішуватися аналітично, однак є випадки коли необхідно знаходити його чисельний розв'язок.

Дискретна оцінка максимальної правдоподібності. Основною ідеєю цього методу є наближення перехідної щільності дискретизованим виразом. Найпростішим прикладом буде розкладання Ейлера-Маруями [73]:

$$X_{i+1} = X_k + \mu(X_i;\theta)\Delta + \sigma(X_i;\theta)\varepsilon_{i,} \quad (3.77)$$

де $\varepsilon_i \sim N(0,\Delta)$– випадкові величини з незалежним нормальним розподілом. Дані оцінки використовуються на практиці навіть не зважаючи на те, що вони виявляються зміщеними в силу спрощення процедури оцінювання.

Останнім часом популярності набуває метод Монте-Карло для марковських ланцюгів (МКМЛ). Основною ідеєю цього методу є рівномірне розбиття інтервалу спостереження даних $[t_i, t_{i+1}]$ на множину менших відрізків за рахунок введення неспостережуваних (згенерованих) величин у проміжні моменти часу [58]. Далі алгоритм оцінювання функціонує таким чином.

1. Генеруються значення X_u^{i+1} з розподілу $P(X_u \mid X_0, \theta^i)$, X_0– спостережувані значення процесу).

2. Генеруються значення параметрів θ^{i+1} з розподілу $P(\theta \mid X_0, X_u{}^i)$.

При досить слабких умовах регулярності згенерований таким чином марковський ланцюг має граничний стаціонарний розподіл $P(X_u, \theta \mid X_0)$, який можна легко перетворити в $P(\theta \mid X_0)$. Основним недоліком методу Монте-Карло

є досить велика ресурсоємність. Однак методи даної групи широко використовуються завдяки універсальності, здатності ураховувати неспостережувані змінні, хорошій масштабованості, незначним похибкам оцінювання та можливості застосування паралельних обчислень для прискорення процесу оцінювання.

В роботі [59] запропоновано застосування для оцінювання моделей гетероскедастичних процесів узагальнений метод моментів. В основі методу - задавання деякої множини умов $\psi_k(X;\theta)$, таких, що $E[\psi_k(X;\theta^*)]=0$, де θ^* – дійсне значення вектора параметрів. Нехай $\psi_{k_i}(X;\theta)=\psi_k(X_i\,|\,\theta)$ тоді оцінку вектора θ можна знайти шляхом мінімізації функціонала $J(\theta)=E[\psi(X;\theta)]^T\sum E[\psi(X;\theta)]$, де $\sum$ – матриця вагових коефіцієнтів, яка була визначена додатно. В роботі [15] було запропоновано використання різних способів побудови цієї матриці. До переваг даного методу можна віднести простоту реалізації, високу швидкодію та узагальненість. Але не враховується вся доступність інформації а також чутливість до похибок дискретизації, які було використано для отримання функцій ψ_k .

Відомий метод непрямого оцінювання, відповідно якому параметри оцінюються не через щільність перехідної ймовірності, а опосередковано за допомогою допоміжної моделі, яка оцінюється за допомогою методому максимальної правдоподібності та забезпечує раціональне наближення до дійсної функції правдоподібності. На даному етапі враховується, що допоміжна модель зазначена невірно, і тому оцінки $\hat{\theta}$ пов'язані з дійсними значеннями θ^* за допомогою функції $\hat{\theta}=\varphi(\theta^*)$, яка визначається шляхом чисельного моделювання. Продуктивність методу і якість одержуваних оцінок сильно залежать від допоміжної моделі. Якщо модель обрана коректно, то навіть при невеликих розмірах вибірки можна отримати високоякісні результати оцінювання. Також запропоновано оцінювати θ шляхом порівняння щільності з її непараметричною оцінкою за спостережуваними даними. Основними недоліками даного методу є те, що реальні дані зазвичай сильно корельовані.

3.5 Оцінювання ринкового ризику на основі рівневих показників

Традиційні міри ризику порівняно погано дозволяють контролювати сам ризик. Ліміти позицій, що визначається факторами ризику або показниками чутливості часто неефективні. Тому на основі цих показників важко застосувати аналіз якості управління портфелем з урахуванням ризику.

Все це пояснює ту велику популярність, якою в сучасному ризик-менеджменті користується підхід до оцінки ризиків на основі показника VaR, який було розглянуто у розділі 1. VaR портфеля для даного довірчого рівня (1-α) і даного періоду підтримки позиції *t*, визначається, як таке значення, яке забезпечує покриття можливих втрат *x* утримувача портфеля за час *t* з імовірністю (1-α) [35, 36]:

$$P(Var \geq x) = 1 - \alpha \quad (3.78)$$

Як випливає із визначення, величина VaR для портфеля заданої структури – це найбільший очікуваний збиток, що спричинений коливанням цін на фінансових ринках, який розраховується:

- на визначений період часу у майбутньому (часовий горизонт);
- із заданою імовірністю його перевищення (рівень довіри);
- при даному припущенні стосовно характеру поведінки ринку (метод розрахунку).

Ключовими моментами для методу VaR є:

- Очікувана величина ризику,яка може бути розрахована в абсолютному вимірі або у процентному відношенні до значення показника на певну дату.
- Часовий горизонт**,** який характеризується очікуваною величиною ризику (іншими словами терміном, за який можна реалізувати на ринку даний підхід без значних втрат).
- Глибина періоду розрахунку VaR – це об’єм ретроспективних або штучно згенерованих даних, на основі яких може бути визначена оцінка.
- Рівень довіри (ймовірність), з якою максимальні збитки не будуть перевищувати розраховану очікувану величину ризику [35].

Тобто формула $P(Var \geq x) = 1 - \alpha$ визначається як очікувана величина ризику VaR перевищує реальну величину ризику x за часовий горизонт t з імовірністю α. (α.=0,01;0,05 і т.ін.).

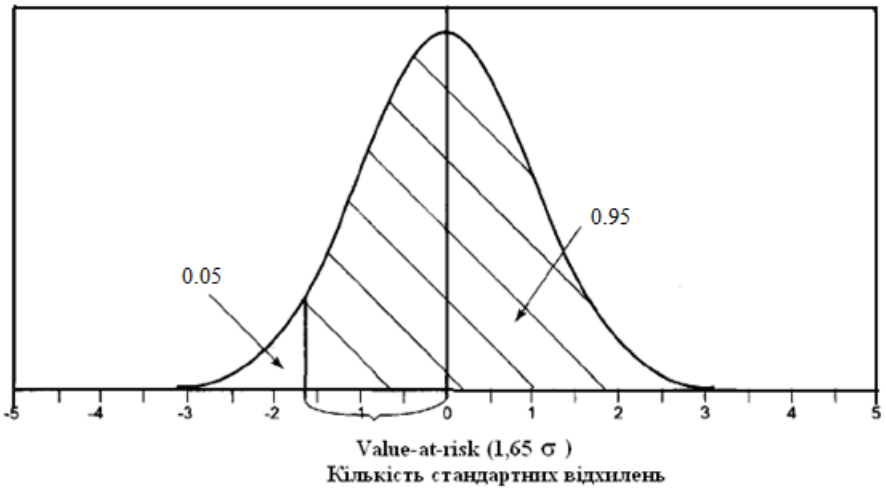

Рис. 3.2 Визначення величини VaR на графіку розподілу прибутків та втрат

Крива на рисунку 3.2 задає розподіл ймовірностей прибутків та збитків для заданих портфеля і періоду підтримки позицій. Заштрихована область відповідає вибраному рівню довіри 95%. VaR представляє собою максимальну величину можливих втрат, що відповідають заданому рівню довіри [35].

Розглянемо непараметричні методи (методи повного оцінювання). Під непараметричними методами будемо розуміти ті методи, в яких оцінювані функції (наприклад, функція розподілу) не визначаються як остаточне число параметрів. Для повної оцінки необхідно провести перерахунок вартості фінансових інструментів без апроксимуючих припущень [35, 36].

Самими поширеними непараметричними методами є:

- метод історичного моделювання;
- метод імітаційного моделювання Монте-Карло.

Ідея цих методів полягає утому, що будується емпірична функція розподілу майбутніх змін цін, і випливаючи з цього потенційних прибутків та втрат [36] .

3.5.1 Оцінювання ринкового ризику на основі непараметричних методів

Історичний варіант VaR-методу розрахунку показника ризикової вартості (VaR) полягає у застосуванні для обчислень реальних історичних значень часового ряду випадкової величини, що аналізується. Розрахунок VaR виконується на основі побудови розподілу для змінених вартостей активів або для цілого портфеля активів за встановлений період часу, та може напряму залежати від одного, або від декількох факторів ризику того ж періоду.

Реалізація методу виконується за такою процедурою:

1) визначаються базову значення (базові показники), які будуть розглянуті для всіх станів ринку, що зафіксовані в історичному періоді;

2) визначаються часові інтервали для розрахунку ризику вартості VaR;

3) визначаються рівні довіри (ймовірність) за якими буде проводитися розрахунок ринкової вартості;

4) розраховуються зміни базових значень з умовою використання базових значень відповідної випадкової величини;

5) упорядковуються за зростанням із формуванням часового ряду змін зміни вартості, які було отримано на попередньому етапі;

6) відокремлюються значення у такому співвідношенні, щоб їх загальна кількість до загальної кількості у часовому ряді змін становило не більше 1-α % для ймовірності α (наприклад, не більше 1 % для ймовірності 99%);

Переваги даного історичного методу моделювання є простота реалізації та відсутність припущень щодо нормального закону розподілу ризику. Також до переваг слід віднести точність у оцінювання ризику нелінійними фінансовими інструментами та відсутність використання помилкової моделі для оцінювання вартості інструмента;

Недоліки методу історичного моделювання є некоректність результатів у випадку, коли вибірка для базового періоду не є репрезентативною по кількості спостережень, а також великий об’єм обчислень для великих диверсифікованих портфелів [36].

3.5.2 Оцінювання ринкового ризику на основі методу імітаційного моделювання Монте-Карло

Метод Монте-Карло (метод стохастичного моделювання), базується на моделюванні випадкових процесів, коли початкові характеристики є заданими. Псевдовипадковим чином генеруються ціни активів до заданих параметрів розподілу, через що кількість сценаріїв може досягати декількох десятків тисяч, а розподіл, що імітується, може бути будь-яким. За іншими характеристиками даний метод аналогічний попередньому методу – методу історичного моделювання.

Моделювання траєкторії цін може відбуватися з використанням різних моделей. Для прикладу розглянемо розповсюджену модель геометричного броуновського руху, яка дає в підсумку вирази для моделювання цін Sна кожному кроці процесу, що складається з дуже великої кількості кроків, та охоплюють період *T:*

$$dS_t = S_t(\mu dt + \sigma dz_t), \tag{3.79}$$

де dz_t–вінеровський випадковий процес. Якщо траєкторія цін складається з n рівних кроків (наприклад, *n* днів), то один крок $\Delta t = \frac{1}{n}$, а випадкова величина εпідпорядкована стандартному нормальному розподілу $(\mu = 0, \sigma = 1)$. Існують й інші моделі еволюції цін, наприклад експоненціальна та інші.

Траєкторія руху цін – це послідовність псевдовипадковим чином змодельованих цін, починаючи від поточної ціни й закінчуючи ціною на деякому кінцевому кроці, наприклад на тисячному або десятитисячному. Чим більше число кроків, тим вища точність методу. Кожна траєкторія являє собою сценарій, за яким визначається ціна на останньому кроці виходячи з поточної ціни. Потім відбувається повна переоцінка портфеля за ціною останнього кроку й розрахунок зміни його вартості для кожного сценарію. Оцінка *VaR* визначається за розподілом змін вартості портфеля [35].

Генерування випадкових чисел у методі Монте-Карло складається із двох кроків. Спочатку можна скористатися генератором випадкових чисел, рівномірно

розподілених на інтервалі між 0 і 1. Потім, використовуючи як аргументи отримані випадкові числа, обчислюють значення функцій розподілів, що моделюються.

Перевагами методу Монте-Карло є висока точність у розрахунках та висока точність у нелінійних цінових характеристиках. Також до переваг слід віднести можливість моделювання історичних і гіпотетичних розподілів з урахуванням ефекту зміни цін. Недоліками методу Монте-Карло вважається складність побудованих моделей та через це – ризик неадекватності цих моделей. Також для розрахунку необхідні великі потужності і витрати часу для проведення експериментів.

3.6 Концепція побудови адаптивної системи для моделювання і прогнозування

На рисунку 3.3 подано структурну схему, що ілюструє системний підхід до організації процесу прогнозування. Він ґрунтується на докладному аналізі досліджуваного процесу, встановленні типів наявних характерних невизначеностей, оцінюванні структури і параметрів моделі та обчисленні оцінок прогнозів за відповідними функціями. Для розв’язання цього комплексу задач необхідно спроектувати і реалізувати СППР. На рисунку 3.4 подано деталізовану схему адаптивного моделювання і прогнозування.

Розглянемо докладніше кожний з етапів побудови СППР. Створення системи для адаптивного моделювання і прогнозування починається з вибору процесу, аналізу поточного стану, існуючих моделей і підходів до його математичного опису та прогнозування подальшого розвитку. Це можуть бути математичні моделі у вигляді систем рівнянь (диференціальних, різницевих або алгебраїчних), закони розподілу вхідних та вихідних величин (статистичні моделі) або логічні моделі у вигляді множин правил, що характеризують логіку взаємодії входів і виходів процесу керування. Вибір типу та структури моделі відіграє суттєву роль для реалізації подальших етапів створення прогнозуючої та керуючої систем.

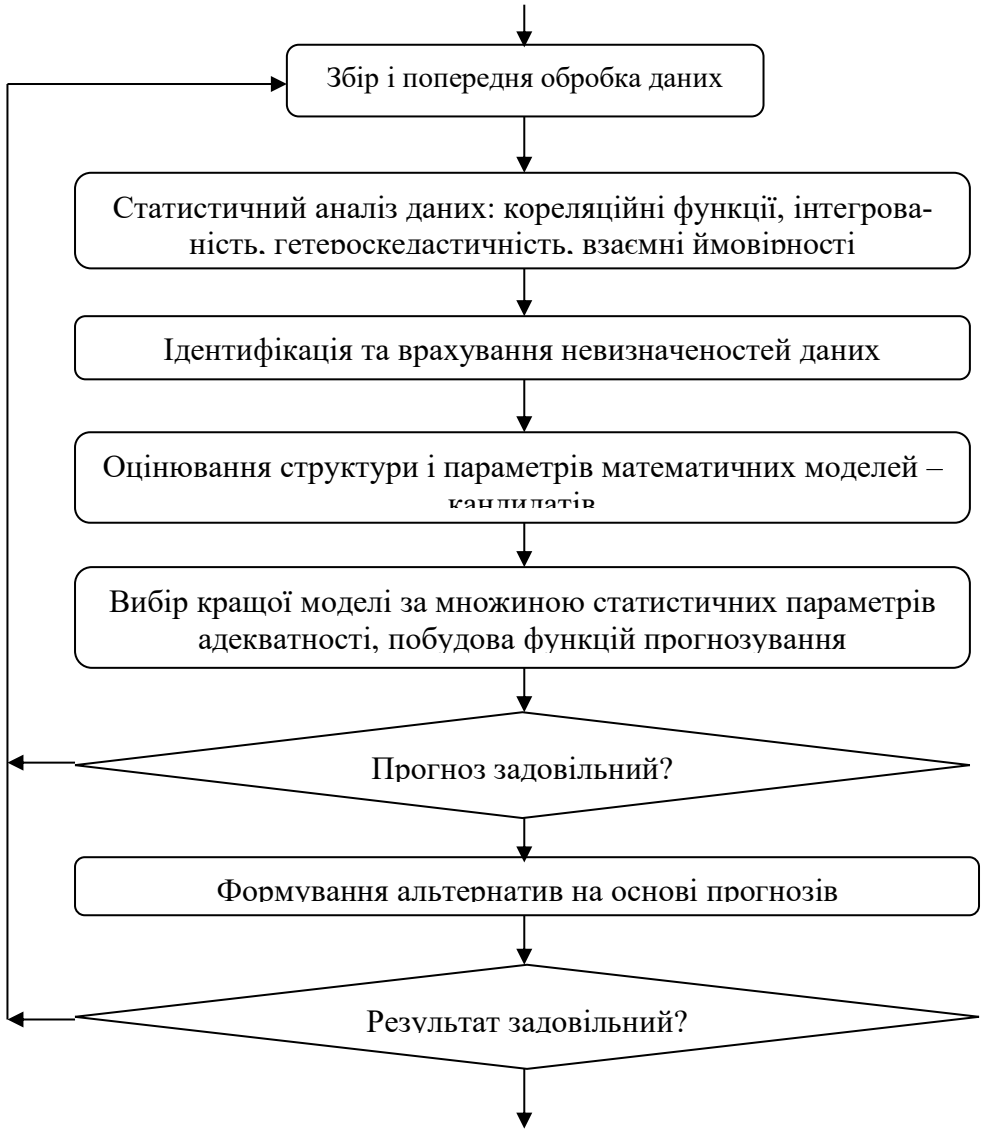

Рис.3.3. Етапи реалізації системного підходу при моделюванні і прогнозуванні

Так модель, створена на основі теоретичних уявлень і закономірностей стосовно конкретного процесу, може потребувати лише деякого подальшого уточнення її параметрів за допомогою статистичних даних. А модель, яка повністю грунтується на статистичних дослідженнях, може потребувати значно більших об’ємів інформації та часу для її побудови. Кожний метод має свої особливості та межі застосування, а тому необхідно знати ці особливості до його практичного застосування.

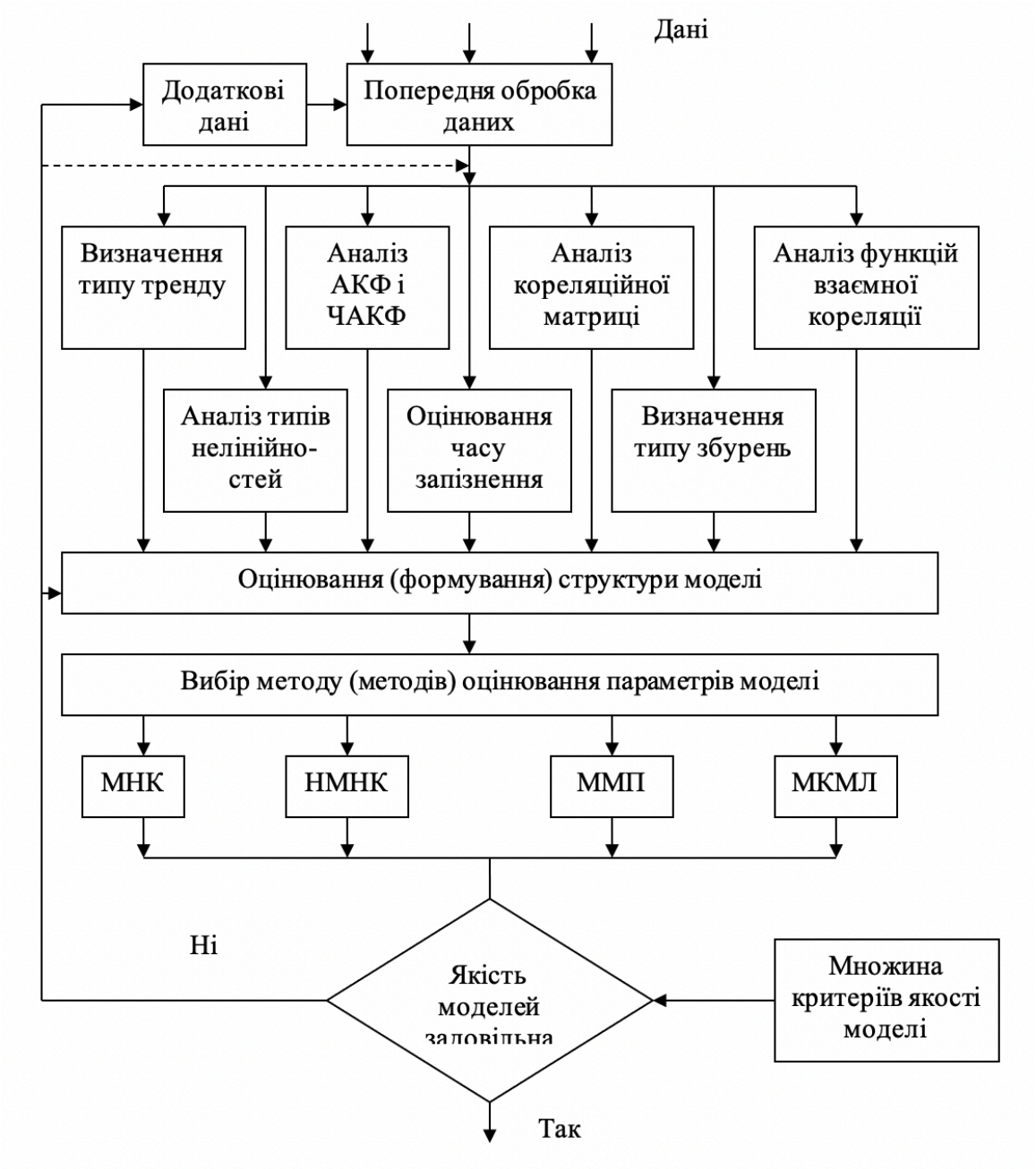

Рис. 3.4. Схема процесу адаптивного оцінювання моделі на основі статистичних (експериментальних) даних

СППР, що створюється, ґрунтується на різновидах регресійних моделей, моделях у просторі станів та байєсівських мережах (ймовірнісні моделі у вигляді спрямованих ациклічних графів). Практика створення прогнозуючих систем для процесів довільної природи свідчить про те, що готові до використання моделі зустрічаються дуже рідко. Навіть існуючі апробовані моделі потребують корегування їх структури та/або параметрів з метою їх адаптування до

конкретних умов використання і даних. Тому у більшості випадків необхідно будувати нову модель на основі поглибленого аналізу процесу та наявних даних. Якість даних відіграє надзвичайно важливу роль для побудови моделі, а тому при формуванні бази даних необхідно керуватись відомими вимогами стосовно їх інформативності, синхронності та коректності.

Попередня обробка даних необхідна для приведення їх до форми, яка забезпечить можливості коректного застосування методів оцінювання параметрів моделі та обчислення їх статистично значущих оцінок. Так, досить часто необхідно застосовувати належне нормування вимірів у заданих межах, заповнювати пропуски даних, корегувати значні імпульсні (екстремальні) значення, логарифмувати великі значення, фільтрувати шумові складові та розв'язувати задачу мультиколінеарності. Фільтрація, розглянута вище, може залежати від конкретної постановки задачі та об'єму наявної інформації про досліджуваний процес (об'єкт).

На основі коректно підготовлених даних оцінюють структури і параметри математичних моделей-кандидатів вибраних процесів. Формування (оцінювання) структури моделі – ключовий момент її побудови. Пропонується визначати структуру моделі такою, що складається із семи елементів:

$$S = \{r,\ p,\ m,\ n,\ d,\ z,\ l\}, \tag{3.80}$$

де r – розмірність моделі (кількість рівнянь, які утворюють модель); p – порядок, тобто максимальний порядок диференціальних або різницевих рівнянь, які формують модель; m – кількість незалежних змінних у правій частині моделі; n – нелінійність та її тип (це нелінійності стосовно змінних або параметрів; також необхідно встановити порядок нелінійності стосовно змінних); d – час затримки (лаг) реакції системи стосовно моменту подачі вхідного впливу та його оцінка; z – зовнішнє збурення процесу та його тип (випадкове або рідко детерміноване); l – можливі обмеження на змінні. Структура оцінюється на основі аналізу особливостей функціонування процесу

та відповідних статистичних даних, які описують його протікання у часі. Як правило, для одного процесу оцінюють кілька моделей-кандидатів, а потім вибирають кращу з них за допомогою множини статистичних параметрів адекватності моделі. Такий підхід суттєво підвищує ймовірність побудови адекватної (кращої) моделі для конкретного застосування.

Часові ряди даних у техніці, економіці та фінансах мають детерміновану і випадкову складові. Поява випадкової складової зумовлена наявністю випадкових збурень, похибок вимірів, неточністю оцінювання структури та обчислень. Тому як статистичну будемо розуміти модель процесу у вигляді розподілу випадкових величин. Обґрунтований вибір типу розподілу та оцінювання його параметрів за допомогою експериментальних даних представляє собою процес побудови статистичної моделі процесу.

Побудована модель, навіть достатньо високого ступеня адекватності, ще не гарантує високої якості оцінок прогнозів, оскільки основна мета побудови прогнозуючої моделі – це належна високоякісна апроксимація основних статистичних характеристик процесу: математичного сподівання, дисперсії та коваріації. Тому після побудови модель необхідно перевірити на можливість її застосування для розв’язання задачі прогнозування. На сьогодні існує широкий спектр методів прогнозування, які застосовують в економіці та фінансах. Однак далеко не всі методи забезпечують високоякісні прогнози у конкретних випадках їх застосування, що зумовлено різноманітними (загаданими вище) причинами, зокрема наявністю невизначеностей та особливостями протікання досліджуваних процесів. Тому вибір методу прогнозування – це не завжди проста задача, яка може потребувати одночасного застосування кількох альтернативних методів і вибору кращого з них на основі аналізу отриманого результату або створення практичних схем оцінювання високоякісних комбінованих прогнозів.

Побудова функцій для багатокрокового прогнозування. Необхідно зазначити деякі корисні особливості моделей авторегресії (АР) та авторегресії з ковзним середнім (АРКС) стосовно обчислення оцінок прогнозів на їх основі. В

узагальненому вигляді таку модель представити таким чином:

$$y(k) = f[y(k-1),...,y(k-p),\, u(k-1),...,u(k-q),\theta] + \varepsilon(k), \qquad (3.81)$$

де $y(k)$ – основна змінна; $u(k)$ – вхідна (керуюча) змінна; p, q – порядок авторегресії та ковзного середнього, відповідно; θ – вектор параметрів моделі; $\varepsilon(k)$ – випадковий процес, поява якого зумовлена наявністю випадкових зовнішніх збурень, похибками вимірів, неточністю структури і параметрів; $k = 0,1,2,...$ – дискретний час, який зв'язаний з неперервним t періодом дискретизації вимірів T_s: $t = k\,T_s$.

Використання в СППР моделей АР і АРКС дає можливість будувати функції прогнозування на основі побудованих моделей. Використання цих функцій спрощує процедури обчислення оцінок багатокрокових прогнозів. За означенням оцінка прогнозу на s кроків стосовно деякого моменту k визначається умовним математичним сподіванням функції, яка дає можливість обчислювати майбутні значення основної змінної за умови, що відома вся необхідна інформація про процес на момент k, включно:

$$\hat{y}(k+s) = E_k[\,y(k+s)\,|\,y(k), y(k-1),..., y(0), \varepsilon(k), \varepsilon(k-1),..., \varepsilon(0)\,], \qquad (3.82)$$

а функція прогнозування, отримана на основі моделі АР(1), має вигляд [9]:

$$\hat{y}(k+s) = E_k[\,y(k+s)] = a_0 \sum_{i=0}^{s-1} a_1^i + a_1^s\, y(k), \qquad (3.83)$$

де $\hat{y}(k+s)$ – оцінка прогнозу змінної $y(k)$ на s кроків; E_k – умовне математичне сподівання стосовно k-го моменту часу; a_0, a_1 – коефіцієнта моделі АР(1). Функція прогнозування процесу АРКС(2,1) на три кроки має вигляд:

$$\hat{y}(k+3) = E_k[y(k+3)] = a_0 + a_1 E_k[y(k+2)] + a_2 E_k[y(k+1)] = \qquad (3.84)$$

$$= a_0(1 + a_1 + a_1^{\,2} + a_2) + (a_1^{\,3} + 2a_1a_2)y(k) + (a_1^{\,2}a_2 + a_2^{\,2})y(k-1) + \\ + \beta_1(a_1^{\,2} + a_2)\varepsilon(k)$$

де a_0, a_1, a_2 і β_1 – коефіцієнти моделі; для обчислення прогнозу використовують $\hat{\varepsilon}(k) = y(k) - \hat{y}(k)$; $y(k)$ – фактичне значення змінної в момент часу k; $\hat{y}(k)$ – оцінка змінної, обчислена за моделлю, тобто $\hat{y}(k) = \theta^T \psi(k)$, де $\psi(k)$ – вектор вимірів змінних у правій частині моделі. Рекурсивна формула для обчислення оцінок прогнозів процесу АРКС(2,1) на довільну кількість кроків s може бути записана так:

$$\hat{y}(k+s) = E_k[y(k+s)] = a_0 + a_1E_k[y(k+s-1)] + a_2E_k[y(k+s-2)]. \qquad (3.85)$$

Такі обчислювальні процедури забезпечують отримання незміщених оцінок прогнозів, дисперсія похибок яких збігається до скінченої константи при збільшенні кількості кроків прогнозування за умови, що: $E[\varepsilon(k)] = 0$ і коваріація $E[\varepsilon(k)\,\varepsilon(j)] = 0$, якщо $k \neq j$.

Одночасне обчислення оптимальних оцінок стану і прогнозів. Задача одночасного оптимального оцінювання стану і прогнозування подальшого руху динамічної системи розв'язується за допомогою згаданих вище методів оптимальної фільтрації, зокрема фільтра Калмана (ФК). На сьогодні існує декілька модифікацій ФК, які забезпечують розв'язання задач оптимального згладжування даних, обчислення оцінок прогнозів за допомогою оптимальних оцінок вектора стану, оцінювання невимірюваних компонент вектора стану процесу та деяких параметрів моделей досліджуваних процесів. Основне рівняння фільтрації для вільної динамічної системи, яке ґрунтується на параметрах моделі процесу у просторі станів, можна записати так:

$$\hat{\mathbf{x}}(k) = \mathbf{A}\,\hat{\mathbf{x}}(k-1) + \mathbf{K}(k)\big[\mathbf{z}(k) - \mathbf{H}\,\mathbf{A}\,\hat{\mathbf{x}}(k-1)\big], \qquad (3.86)$$

де $\hat{\mathbf{x}}(k)$ – оптимальна оцінка вектора стану $\mathbf{x}(k)$ у момент часу k; $\mathbf{A}$ – перехідна матриця станів процесу; $\mathbf{z}(k)$ – вектор вимірів змінних на виході

об’єкта; **H** – матриця (коефіцієнтів) вимірів; $\mathbf{K}(k)$ – оптимальний матричний коефіцієнт фільтра, який обчислюється за умови мінімізації функціоналу:

$$J = \min_{\mathbf{K}} E\left\{[\hat{\mathbf{x}}(k) - \mathbf{x}(k)]^T\ [\hat{\mathbf{x}}(k) - \mathbf{x}(k)]\right\}, \tag{3.87}$$

тобто за умови мінімуму математичного сподівання суми квадратів похибок оцінок вектора стану процесу (значення **K** визначається розв’язком відповідного рівняння Ріккаті). Алгоритм оцінювання вектора стану формує також однокроковий прогноз вектора стану:

$$\hat{\mathbf{x}}(k+1,k) = \mathbf{A}\,\hat{\mathbf{x}}(k), \tag{3.88}$$

за допомогою якого можна отримати оцінки прогнозів на довільну кількість кроків s:

$$\hat{\mathbf{x}}(k+s,k) = \mathbf{A}^s\ \hat{\mathbf{x}}(k). \tag{3.89}$$

Таким чином, цінність фільтра полягає у тому, що він виконує роль пристрою для згладжування і прогнозування, а тому його введення в СППР надає системі додаткові корисні функціональні можливості. Крім того, адаптивний фільтр дає можливість оцінювати статистичні характеристики збурення стану і похибок вимірів, які не завжди можна визначити апріорно [7, 8, 12].

Адаптація байєсівської мережі. Одним із потужних сучасних імовірнісних інструментів розв’язання задач прогнозування, класифікації та підтримки прийняття рішень є графічні моделі причинно-наслідкових зв’язків у формі байєсівських мереж (БМ) [10, 15]. Для побудови структури БМ використано алгоритм на основі статистичного аналізу рядів даних, які характеризують еволюцію змінних мережі [10]. В СППР реалізовано **алгоритм адаптування** структури мережі до нових даних, що надходять в реальному часі. Для пояснення процедури адаптації мережі введемо такі позначення: $Z = \{X_1,...,X_n\}$ – множина вузлів БМ, яка визначається числом змінних в базі даних; $E = \{(X_i, X_j) | X_i, X_j \in Z\}$ – множина дуг мережі; X_i – вузол БМ, що

відповідає спостереженням однієї змінної з бази даних; $n = |Z|$ – число вузлів БМ; r_i – число значень, що можуть прийматися вузлом X_i; v_{ik} – k-е значення змінної X_i; Π_i – множина вузлів-предків вузла X_i; ϕ_i – множина можливих ініціалізацій Π_i; $q_i = |\phi_i|$ – число можливих ініціалізацій Π_i; ϕ_{ij} – j-а ініціалізація множини вузлів-предків Π_i вузла X_i; B_S – структура БМ; B_P – імовірнісна специфікація БМ, тобто частина опису моделі, що представляє імовірнісні характеристики БМ; $\theta_{ijk} = p(X_i = v_{ik} \mid \phi_{ij}, B_P)$ при цьому сума ймовірностей $\sum_k \theta_{ijk} = 1$; $f(\theta_{ij1},...,\theta_{ijr_i})$ – щільність розподілу імовірностей для вузла X_i та ініціалізації ϕ_{ij}; D_0 – вихідна база даних спостережень; S_0 – структура БМ, отримана внаслідок попередньої пакетної обробки бази D_0; D_1 – база даних нових спостережень, не використаних при побудові S_0; S_1 – структура БМ, отримана після адаптації S_0 до нових даних D_1. Ставилась задача розробки алгоритму адаптування вихідної байєсівської мережі $G = \langle Z, E \rangle$ із структурою S_0, побудовану за вихідною базою спостережень D_0, до нових спостережень D_1. Тобто необхідно сформувати оновлену структуру мережі $S_1 \Leftrightarrow D_1$. При цьому експериментальні (статистичні) дані можуть мати довільний розподіл імовірностей, а процеси, які описуються цими даними, можуть мати нестаціонарний характер, тобто математичне сподівання $M[X_i] \neq const$ і дисперсія $M\{X_i - M[X_i]\}^2 \neq const$.

Адаптація побудованої мережі до нових даних виконується у такій послідовності:

1. Реалізація процедури корегування структурної частини моделі:
   – процедура видалення дуг, що не відповідають множині даних;
   – додавання нових дуг.
2. Процедура корегування імовірнісної частини моделі.

Оскільки на початковому етапі навчання БМ імовірнісну складову моделі представляють таблиці умовних розподілів ймовірностей (ТУЙ), отримані безпосередньо на підставі частотного аналізу появи значень змінних у спостереженнях, тому відразу визначимо зміни в процедурі корегування імовірнісної частини моделі. З метою полегшення проведення корегування імовірнісної частини моделі, корисно зберігати не таблиці розподілу умовних імовірностей, а значення N_{ijk}. Це дає можливість швидше оновлювати дані щодо розподілу умовних імовірностей, а самі значення умовних імовірностей можна буде обчислити, користуючись формулою Діріхле:

$$p(X_i = v_{ik} \mid \Pi_i = \phi_{ij}) = \frac{N_{ijk} + 1}{N_{ij} + r_i}. \tag{3.90}$$

При корегуванні структури БМ порядок обходу вузлів визначаємо за вкладом кожного вузла в значення умовної ймовірності $p(D_1 \mid D_0, S_0) = \prod_{i=1}^{n} \frac{\prod_{s=1}^{R_i} \prod_{t=1}^{Q_i} \prod_{u=1}^{m_{its}} (N_{its} + u)}{\prod_{t=1}^{Q_i} \prod_{u=1}^{M_{it}} (N_{it} + r_i - 1 + u)}$. Суть аналізу інформаційної важливості дуг полягає у такому. На етапі перевірки дуг на необхідність видалення для кожного вузла обчислюється значення $K_{delete}(S_0)$ для поточної конфігурації множини вузлів-предків, а також значення $K_{delete}(S_{-1}^m)$ для конфігурацій, які представляють собою результат видалення однієї з M $(1 \le m \le M)$ вхідних дуг з поточного вузла. Якщо виконується умова $K_{delete}(S_{-1}^m) \le K_{delete}(S_0)$, то m-а дуга залишається в структурі мережі, оскільки видалення даної дуги призводить до зменшення значення локального функціоналу якості (тобто для поточного вузла). Інакше дуга заноситься в список дуг, що підлягають подальшій перевірці на необхідність видалення. Список відсортовується за збільшенням значення $K_{delete}(S_{-1}^m)$. Список (множина) дуг аналізується послідовно. Подальша перевірка полягає у обчисленні значення локального функціоналу якості при вихідній конфігурації і конфігураціях, які ми отримуємо при видаленні однієї з дуг, що

залишилися в списку. Тактика вилучення і додавання дуг застосована у інкрементному варіанті адаптаційного алгоритму, наведеному нижче. Оскільки результатом реалізації байєсівського підходу є вибір стратегії адаптації на основі функціоналу:

$$P(S_1 \mid D_1, D_0, S_0) = \arg\max_S \frac{P(S \mid D_0) P(D_1 \mid S, D_0)}{P(D_1 \mid S_0, D_0)}, \qquad (3.91)$$

то процедура вилучення і додавання дуг здійснюється таким чином.

Якщо врахувати вид розв'язку оптимізаційної задачі адаптації БМ, то тактика вилучення дуг повинна приводити до зменшення першої складової чисельника $P(S \mid D_0)$, оскільки вона досягає максимуму при $S = S_0$ в результаті формування початкової структури БМ. Таким чином, для отримання позитивного ефекту від адаптації необхідно компенсувати втрати від вилучення дуги ефектом від додавання нової дуги. Оскільки вихідною умовою алгоритму К2 є наявність впорядкованої послідовності вузлів, то пошук дуги-претендента на додавання здійснюється саме в такому порядку. Оцінка дуги виконується шляхом обчислення значення локального функціоналу якості. Відповідно, претендент на додавання повинен визначати конфігурацію вхідних дуг, що має найбільше значення локального функціоналу якості [9].

Оцінювання адекватності моделей і якості прогнозів. Важливим моментом процесу прогнозування є об'єктивне визначення якості отриманого прогнозу. Оскільки оцінки прогнозів – це випадкові величини, то для визначення їх якості необхідно використовувати множину відповідних статистичних критеріїв. Саме множину, а не один критерій, оскільки кожен критерій характеризує одну властивість оцінки прогнозу. Іноді якість оцінок прогнозів визначають лише за допомогою середньоквадратичної похибки (СКП). Однак, значення СКП – це лише одна із множини можливих статистик, яка залежить від масштабу даних, а тому цієї характеристики явно недостатньо для аналізу якості прогнозу.

Якість лінійних та псевдолінійних моделей оцінюють за допомогою декількох статистичних критеріїв якості, зокрема таких: коефіцієнт множинної

детермінації (R^2), який характеризує інформативність моделі по відношенню до інформативності даних; статистика Дарбіна-Уотсона (DW), що визначає ступінь автокорельованості похибок моделі; інформаційний критерій Акайке (AIC) і статистика Байєса-Шварца (BSC); сума квадратів похибок моделі ($\sum e^2(k)$); F – статистика Фішера та інші.

Поглиблене оцінювання якості прогнозів досягається за рахунок використання критеріїв, які дають відносні оцінки якості (наприклад, коефіцієнт Тейла) та оцінки якості у процентах (наприклад, середня абсолютна похибка у процентах ($САПП$)). Переваги їх використання полягають у тому, що вони не залежать від масштабу даних і легко інтерпретуються ОПР. $САПП$ і коефіцієнт Тейла обчислюють за виразами:

$$САПП = \frac{1}{s}\sum_{i=1}^{s}\frac{|y(k+i)-\hat{y}(k+i,k)|}{|y(k+i)|}\times 100\% = \frac{1}{s}\sum_{i=1}^{s}\frac{|e(k+i)|}{|y(k+i)|}\times 100\%,$$

$$U = \frac{\sqrt{\frac{1}{s}\sum_{k=1}^{s}[y(k+i)-\hat{y}(k+i)]^2}}{\sqrt{\frac{1}{s}\sum_{i=1}^{s}y^2(k+i)}+\sqrt{\frac{1}{s}\sum_{i=1}^{s}\hat{y}^2(k+i)}}, \qquad (3.92)$$

де s - кількість кроків прогнозування; $y(k+i)-$ фактичні значення даних; $\hat{y}(k+i)-$ оцінки прогнозів відносно $k-$го моменту часу, на який наявна вся інформація про досліджуваний процес. Коефіцієнт Тейла U – це важлива характеристика якості моделі і прогнозу, за означенням $0 \le U \le 1$. Якщо $U \to 0$, то оцінки прогнозів наближаються до фактичних значень ряду і модель має високу ступінь адекватності. Тобто U дає можливість встановити придатність моделі (і методу оцінювання прогнозу на її основі) для оцінювання прогнозу в принципі.

Для автоматизованого вибору кращої моделі можна скористатись інтегральним критерієм якості [9]:

$$V_N(\theta, D_N) = e^{1-R^2} + \frac{SSE}{N} + \begin{cases} \ln(AIC + BSC), & якщо \quad AIC + BSC > 0 \\ e^{AIC+BSC}, & якщо \quad AIC + BSC \le 0 \end{cases} + e^{2-DW} + \\ + \ln(СКП) + \ln(САПП) + e^{U} \qquad (3.93)$$

де D_N – дані, що використовуються для оцінювання структури і параметрів моделі; $СКП$ – середньоквадратична похибка однокрокового прогнозу на навчальній (історичній) вибірці; $САПП$ – середня абсолютна похибка прогнозу в процентах; U – коефіцієнт Тейла (наближається до нуля, якщо модель придатна для прогнозування).

Альтернативним варіантом використаного інтегрального критерію є такий:

$$V_N(\theta, D_N) = e^{|1-R^2|} + \ln(1 + \frac{SSE}{N}) + e^{|2-DW|} + \ln(1 + СКП) + \ln(1 + САПП) + e^{U} \qquad (3.94)$$

де N – кількість вимірів часового ряду даних. Потужність цих критеріїв перевірено експериментально і встановлено, що вони дають можливість вибрати кращу модель практично з одиничною ймовірністю.

У багатьох випадках кращих результатів прогнозування можна досягти за рахунок усереднення або комбінування за допомогою вагових коефіцієнтів оцінок прогнозів, отриманих за допомогою різних методів. При цьому необхідно задовольнити такі умови: похибки оцінок прогнозів, отриманих за різними методами, мають бути некорельованими, а дисперсії цих похибок близькими за своїми значеннями.

Адаптивне обчислення оцінок прогнозів. Для збереження якості оцінок прогнозів в умовах нестаціонарності досліджуваного процесу, а також для підвищення якості прогнозування процесів з довільними статистичними характеристиками необхідно застосовувати адаптивні схеми оцінювання прогнозів. Вихідними величинами для аналізу якості прогнозів та формування адаптивних схем їх оцінювання є значення похибок прогнозів та статистичні характеристики їх якості. Для розв'язання задачі *структурної адаптації* прогнозуючої моделі до змін у досліджуваному процесі та до вимог стосовно

якості прогнозу можна скористатись такими обчислювальними можливостями:

– періодичний аналіз типу розподілу даних і його параметрів та врахування отриманого результату при виборі методу оцінювання параметрів моделі;

– автоматизований аналіз часткової автокореляційної функції (ЧАКФ) залежної (основної) змінної з подальшим корегуванням структури моделі шляхом введення/вилучення додаткових лагових значень;

– почергове введення у модель можливих регресорів та аналіз їх впливу на якість прогнозу; особливо корисними для оцінювання прогнозів є регресори, які вводяться в модель з лагами більшими одиниці – це так звані *провідні індикатори*, що надають можливість коректно обчислювати прогнози на ту кількість кроків, що відповідає фактичному лагу; формування додаткових індикаторів на основі регресорів;

– автоматизований аналіз функції часткової взаємної кореляції основної змінної з регресорами з метою корегування лагових значень регресора у правій частині рівняння;

– автоматизований вибір оптимальних вагових коефіцієнтів в процедурах експоненційного згладжування, пошуку подібних траєкторій, регресії на опорних векторах та деяких інших методах;

– автоматизований аналіз залишків регресійних моделей з метою встановлення їх інформативності та корегування структури моделі процесу на основі результатів аналізу;

– адаптивне формування масивів вимірів змінних стану процесу за допомогою методів ієрархічного комплексування (інтегрування) даних, що забезпечує підвищення їх інформативності.

Задача *параметричної адаптації* моделі до даних розв’язується завдяки застосуванню повторного (рекурсивного) оцінювання параметрів математичних і статистичних моделей з надходженням нових даних, що сприяє уточненню параметрів моделі та підвищенню якості прогнозу. При цьому для оцінювання однакових структур застосовуються різні методи, що надає можливість отримання додаткових моделей-кандидатів для подальшого аналізу.

Застосування тієї чи іншої схеми адаптації залежить від конкретної постановки задачі, якості та наявного об’єму експериментальних (статистичних) даних, сформульованих вимог до якості оцінок прогнозів та часу, який може бути наданий для виконання обчислень. Кожний метод адаптивного формування оцінки прогнозу має свої особливості, які мають бути враховані при створенні системи адаптивного прогнозування.

3.7 Висновки до розділу

Розглянуто процедури підготовки (фільтрації) даних до моделювання випадкових процесів, що містять детерміновану складову. Показано, що методологія побудови математичних моделей нелінійних нестаціонарних процесів включає в себе кілька етапів. Дотримування існуючих методик та застосовування статистичних параметрів якості на кожному із етапів дозволить уникнути неадекватності моделей. Це дає можливість досягти високої якості проміжних та остаточних результатів обчислювальних експериментів. Найбільш важливими є у даному випадку такі принципи: – ієрархічність процесу моделювання і прогнозування; – адаптивність процедур оцінювання структури і параметрів моделей; – ідентифікація та врахування можливих невизначеностей; використання адаптивних процедур моделювання.

Процес попередньої обробки даних за допомогою фільтрації є дуже важливим етапом аналізу даних. Зазвичай, застосування методів на цьому етапі, дає можливість значно покращити результати досліджень. Інколи відсутність методів попередньої обробки ставить під загрозу всі подальші кроки по обробці даних та оцінюванні ризиків. Це може призводити до низької якості результатів, наприклад, оцінки прогнозів характеризуються великими похибками. Адаптивне прогнозування нелінійних нестаціонарних процесів є також однією з ключових задач сучасності, у зв’язку з тим, що більшість процесів в економіці, фінансах, екології та технологічних процесах дуже швидко змінюються та не мають єдиного підходу.

Розглянуто задачу тестування на наявність гетероскедастичності та структури деяких моделей гетероскедастичних процесів, що мають широке розповсюдження у фінансах. Показано, що моделювання таких процесів вимагає створення двох моделей – для самого процесу та його умовної дисперсії, яка використовується надалі для оцінювання ризику. Наприклад, прогноз дисперсії використовується для оцінювання ринкового ризику, одного із самих розповсюджених видів ризику у фінансах.

РОЗДІЛ 4
ПРОВЕДЕННЯ ОБЧИСЛЮВАЛЬНИХ ЕКСПЕРИМЕНТІВ

4.1 Розробка критеріальної бази для аналізу якості результат

4.1.1. Початкові дані для статистичного аналізу та прогнозування

Прогнозування полягає в заснованому на відповідному статистичному аналізі формальному описі стану досліджуваної системи або процесу через один, два або більше тактів часу по відношенню до поточного моменту часу, тобто до сьогодення. Оцінки прогнозів мають властивість наукового результату. Іншими словами, в основі прогнозу лежить наукове обґрунтування, яке може бути відтворене і без автора прогнозу. Експертна оцінка, тобто прогноз фахівця в даній конкретній області, являє собою певний проміжний варіант підходу до формування уявлення про майбутнє. Оскільки одного боку, ця оцінка заснована на суб'єктивному представленні експерта про можливий розвиток прогнозованого процесу, а з іншого, – вона враховує багато факторів, якщо не піддаються безпосередньому вимірюванню і формалізації, то допускається об'єктивна інтерпретація в рамках наукового обґрунтування експерта [21].

Статистичні методи аналізу та прогнозування засновані зазвичай на глибокій обробці статистичних даних, що відносяться до досліджуваного процесу. При цьому існують такі особливості виконання такого дослідження.

1. Основні джерела вихідних статистичних даних ділять на первинні і вторинні.

До первинних джерел відносять спеціальні вибіркові обстеження, опитування, переписи, спрямовані на отримання тих даних і в такій формі, які необхідні саме для запланованих прогнозних розрахунків або управлінських рішень. Отримання вихідних статистичних даних з первинних джерел пов'язано зі спеціально спланованою роботою (і, відповідно, з виділенням для цього спеціальних засобів). Планується склад показників (процесів), спосіб організації вибірки, а іноді і фіксовані значення деяких показників, при яких проводиться реєстрація значень інших показників [22].

Вторинні джерела – це опубліковані в тому чи іншому вигляді вихідні дані, вже зібрані кимось поза прямим зв'язком з конкретним завданням прогнозиста, але надають інформацію, в тій чи іншій мірі корисну саме для розв’язання цієї конкретної задачі.

2. Вимоги, що пред'являються до вихідних статистичних даних. Формуючи масив вихідних статистичних даних з первинних або вторинних джерел, необхідно притримуватись основних вимог до якості цих даних.

Релевантність. Це властивість означає, що використовувані дані (тобто обрані для аналізу змінні, методологія і час їх виміру) повинні відображати саме аналіз діяльності і повинні бути «прив'язані» до потрібних об'єктів і відповідних моментів часу.

Надійність і точність. Це властивість вихідних даних досягається за допомогою різних (прямих і непрямих) методів перевірки надійності використовуваних джерел, дотримання прийнятої методології вимірювань, достовірності відповідей респондентів, встановлення можливих збоїв і помилок в їх записах.

Порівнянність. Самі дані повинні супроводжуватися такими коментарями та поясненнями, що стосуються змісту аналізованих показників і методології їх вимірювання, які дозволили б зберегти можливість їх зіставлення (в часі і просторі) і «приведення до спільного знаменника» в ситуаціях, що характеризуються змінами в методології вимірювань і коригуванням складу аналізованих змінних [22].

Репрезентативність. Дотримання цієї властивості досягається таким способом організації вибірки, при якому вона повно і адекватно передає досліджувані властивості всієї аналізованої сукупності (тобто тієї сукупності, від якої ця вибірка відбиралася).

Обчислення прогнозу і виконання пов'язаних з ним побудови і експериментальної перевірки (верифікація) ймовірнісно-статистичної моделі зазвичай засновані на одночасному використанні інформації двох типів:

- апріорної інформації про природу і змістовну сутність аналізованого явища, представленої, як правило, у вигляді тих чи інших теоретичних закономірностей, обмежень, гіпотез;
- вихідних статистичних даних, що характеризують процес і результати функціонування аналізованого явища або системи.

Можна виділити такі основні етапи прогнозування.

1-й етап (постановка завдання) включає в себе визначення кінцевих прикладних цілей прогнозування; набор чинників і показників (змінних), опис взаємозв'язків між атрибутами, що нас цікавлять; ролі цих чинників і показників – які з них, в рамках поставленої конкретної задачі, можна вважати вхідними (тобто повністю або частково регульованими або хоча б легко піддаються реєстрації і прогнозуванню; подібні фактори несуть смислове навантаження (пояснюються моделлю), а які – вихідними [23].

2-й етап (апріорний, перед модельний аналіз) базується на попередньому аналізі змістовної сутності досліджуваного процесу або явища. На цьому етапі формуються та формалізуються апріорні знання про ці явища у вигляді ряду гіпотез та припущень.

3-й етап (інформаційно-статистичний) полягає у зборі необхідної статистичної інформації, тобто реєстрації значень, що беруть участь в аналізі чинників і показників на різних часових і (або) просторових тактах функціонування модельованої системи.

4-й етап: специфікація моделі, яка спирається на прийняті на 2-му етапі гіпотези і початкові допущення. Вона включає в себе формування загального вигляду модельних співвідношень, що зв'язують між собою вхідні і вихідні змінні. Говорячи про загальний вигляд модельних співвідношень, мається на увазі та обставина, що на даному етапі буде визначена лише структура моделі, її аналітичний запис, в якій поряд з відомими числовими значеннями (представленими в основному вихідними статистичними даними) будуть наявні величини, змістовний сенс яких визначено, а числових значень ще немає (їх зазвичай називають параметрами моделі, невідомі значення яких підлягають

статистичному оцінюванню).

5-й етап: ідентифікація моделі, складається в проведенні статистичного аналізу даних з метою «налаштування» значень невідомих параметрів на ті вихідні статистичні дані, які ми маємо. В процесі реалізації цього етапу необхідно спочатку відповісти на запитання, чи можливо в принципі відновити значення невідомих параметрів моделі за наявними вихідними статистичними даними для прийнятої на 4-му етапі структури моделі. Це становить так звану проблему ідентифікації моделі. Після позитивної відповіді на це питання необхідно вирішити вже проблему параметричної ідентифікації моделі, тобто запропонувати і реалізувати математично коректну процедуру оцінювання невідомих значень параметрів моделі за наявними вихідними статистичними даними. Якщо проблему ідентифікації вирішити неможливо, то повертаємося до 4-го етапу і вносимо необхідні корективи в розв'зання задачі специфікації моделі [23].

6-й етап: верифікація моделі, яка полягає у використанні різних процедур порівняння висновків, отриманих за моделлю, оцінок, наслідків з дійсністю. Цей етап також відомий як статистичний аналіз точності та адекватності моделі. Якщо результати даного етапу песимістичні, тоді необхідно повернутися до етапу 4, а в деяких випадках до етапу 1. У випадку, коли даний етап дає позитивні результати, модель може використовуватися для оцінювання прогнозу відповідно до описаної вище загальної схеми [22, 23].

В описі змісту 1-го етапу процедури прогнозування вказувалось, зокрема, про необхідність визначення кінцевих прикладних цілей прогнозування. Це має на увазі, зокрема, і визначення необхідного типу прогнозу. Тип прогнозу визначається двома факторами:

- горизонтом прогнозування;
- ієрархічним рівнем прогнозованого показника.

За горизонтом прогнозування прогнози діляться на короткострокові (на 1-2 такту часу вперед), середньострокові (на 3-5 тактів) і довгострокові (більше ніж на 5 тактів часу вперед). Так часу визначається, як правило, періодом дискретизації вимірів.

Математична модель (ММ). Математична модель – це математична конструкція, що представляє собою абстракцію реального світу: в моделі дослідника цікавлять відносини між реальними елементами, замінені відповідними відносинами між елементами математичної конструкції (математичними категоріями). Ці відносини, як правило, представлені у формі рівнянь і/або нерівностей між змінними, що характеризують функціонування модельованої реальної системи. Складність побудови математичної моделі полягає у тому, щоб поєднати якомога більшу лаконічність в її математичному описі з достатньою точністю модельного відтворення саме тих сторін аналізованої реальності, які цікавлять дослідника.

4.1.2 Вимоги до математичної моделі

При побудові ММ використовують різні математичні засоби опису об'єкта – теорію множин, теорію графів, теорію ймовірностей, математичну логіку, математичне програмування, диференціальні або інтегральні рівняння іт. ін. Загальними вимогами до моделі є такі:

1. Модель повинна буди адекватною процесу чи об‘єкту.

Адекватність означає, що модель повинна:

а) відображати найбільш характерні зв‘язки та взаємодію між змінними процесу;

б) враховувати можливі керуючі дії (сигнали); в) враховувати вплив зовнішніх збурень та шуми вимірювань;

в) враховувати початкові значення змінних та обмеження на них.

Формально адекватність визначають за допомогою ряду статистичних величин [25]. Наприклад, дуже часто використовують середньо-квадратичну похибку моделі (СКП), яка обчислюється за формулою:

$$\text{СКП}(x_s, x_m) = \sqrt{\frac{1}{n}\sum_{k=1}^{N}[x_s(k) - x_m(k)]^2}\,, \qquad (4.1)$$

де $x_s(k) -$ вимір вихідного сигналу об′єкту в момент k;

$x_m(k)$ −оцінка виміру за моделлю.

Використання одного параметра для визначення ступеня адекватності моделі є некоректним підходом, оскільки оцінки параметрів – це випадкові величини,а тому збільшення кількості критеріїв адекватності сприяє підвищенню ймовірності вибору адекватної (кращої) моделі.

2. Кожне рівняння моделі повинно мати хоча б один аналітичний розв‘язок або, у випадку коли це неможливо,то чисельний розв‘язок. Розв‘язок необхідний для аналізу поведінки процесу (аналізу збіжності) та для обчислення оцінок прогнозів [10].

Одним із принципів, яких необхідно дотримуватись при побудові моделі є такий: в моделі не повинно бути нічого зайвого крім необхідного. Звичайно,що дотримуватись цього принципу досить непросто і на практиці буває так, що модель дійсно має надзвичайно складну структуру, що також може бути оправдано необхідністю досягнення високого ступеня її адекватності процесу. Це особливо стосується нелінійних процесів [6]. Але при побудові лінійних моделей у вигляді авто регресії (АР) чи авторегресії з ковзним середнім (АРКС) достатньо побудувати модель, статистичні характеристики якої співпадають з статистичними характеристиками часового ряду, на основі якого вона оцінюється. Такі спрощені моделі виявляються цілком придатними для прогнозування та керування процесами. Загалом питання складності моделі вирішується в кожному випадку окремо.

3. Модель повинна буди достатньо універсальною для того щоб її можна було застосувати до описання класу однотипових процесів або до описання функціонування процесу в різних умовах.

Наприклад, для описання моторної функції людини (реакція на зовнішні збуджуючі сигнали) застосовують звичайне диференціальне рівняння другого порядку, яке представляють у вигляді функції передачі такого ж порядку:

$$W(s) = \frac{Ke^{-\tau s}}{(1-T_1 s)(1-T_2 s)}, \qquad (4.2)$$

де К – статичний коефіцієнт передачі об’єкта;τ – час запізнення по входу; який в середньому дорівнює для людини 300-350 мс;T_1, T_2 – постійні часу [5, 6].

Така передаточна функція може використовуватись, наприклад, для

описання реакції людини на зовнішні відео- або аудіосигнали, що поступають через систему візуального сприйняття чи аудіосистему (поширений приклад – водіння автомобіля або управління іншою машиною). Значення параметрів моделі можуть бути різними для різних людей, але структура моделі залишається незмінною. Таким чином, модель (4.2) описує широкий клас біологічних систем і цілком відповідає умові універсальності [26].

При моделюванні технічних систем широко застосовують ланки першого і другого порядку, що відповідають звичайним диференціальним рівнянням таких же порядків. На основі таких простих ланок можна побудувати моделі будь-якої складності. Дуже поширений в техніці та екології клас систем з розподіленими параметрами. Наприклад, процес розповсюдження домішок в атмосфері та водному середовищі, механічні коливання сонячних батарей та антен супутників, крила літака, локомотива з вагонами на залізниці, автомобіля з причепом і багато інших. Динаміку таких систем описують диференціальними рівняннями з частинними похідними.

4. Вимога робастності (robust – сильний, міцний). Робастність означає, що модель повинна давати прийнятний прогноз вихідної змінної не тільки на тому відрізку часового ряду, на основі якого вона побудована, але і на будь-якому іншому відрізку, що відповідає вибраному режиму роботи. Робастність може розглядатись також як стійкість моделі по відношенню до збурень, похибок та пропусків вимірів. Вимога робастності є особливо критичною для систем, що працюють в реальному часі, оскільки нестійка модель може стати причиною створення аварійної ситуації [4, 5].

5. Вимога адаптивності. Ця вимога означає, що хоча б частину параметрів моделі (щонайменше один) можна уточнювати по мірі надходження нових вимірів від об‘єкта. Ця вимога є обов‘язковою при побудові моделей нестаціонарних систем, тобто систем, параметри яких є функціями часу. Системи керування, побудовані для нестаціонарних процесів, називають адаптивними. Такі системи є досить складними з точки зору аналізу збіжності оцінок параметрів та похибок керування, а тому при проектуванні адаптивних

систем необхідно особливу увагу приділяти питанням достатнього збудження процесу, вибору методу оцінювання параметрів, збіжності оцінок параметрів, оцінювання прогнозів та значень керуючих впливів. [26]

4.1.3 Аналіз якості моделі

Адекватність регресійних моделей може бути встановлена на основі аналізу послідовності залишків (похибок моделі); при цьому розрахункові значення знаходять підстановкою в модель фактичних значень всіх включених у модель факторів. Залишкова послідовність перевіряється на виконання властивостей випадкової компоненти економічного часового ряду: близькість нулю математичного очікування, випадковий характер відхилень, відсутність автокореляції та нормальність закону розподілу.

Аналіз адекватності моделі виконується на основі відповідних параметрів адекватності, до яких відносяться подані нижче.

- Сума квадратів залишків – сума квадратів величин розбіжності між змодельованими і фактичними значеннями пояснюючої змінної на період і ідентифікації, яка розраховується за такою формулою:

$$R^2 = \sum_{i=1}^{n} (y_i - \hat{y_i})^2 \tag{4.3}$$

- Коефіцієнт детермінації — статистичний параметр, щовикористовуєтьсявстатистичнихмоделяхякміраінформативностімоделістосов ноданих. Він показує на скільки наявні спостереження відповідають моделі:

$$R^2 = 1 - \frac{V(y|x)}{V(y)} = 1 - \frac{\sigma^2}{\sigma_y^2} \tag{4.4}$$

де $V(y|x) = \sigma^2$ – дисперсія основної змінної, оціненої за моделлю.

- Критерій Дарбіна-Уотсона (або DW критерій) – статистичний критерій, який використовується для тестування автокореляції першого порядку елементів досліджуваної послідовності. Найбільш часто застосовується при аналізі часових рядів і залишків регресійних моделей:

$$d = \frac{\sum_{t=2}^{n}(e_t - e_{t-1})^2}{\sum_{t=1}^{n} e_t^2} \approx 2(1 - \rho_1) \quad (4.5)$$

де ρ_1 - коефіцієнт автокореляції першого порядку. Якщо автокореляція відсутня $d = 2$, в іному випадку d прямує до нуля, а при негативній дорівнює 4:

$$\begin{cases} \rho_1 = 0 \ \rightarrow d = 2 \\ \rho_1 = 1 \ \rightarrow d = 0 \\ \rho_1 = -1 \ \rightarrow d = 4 \end{cases} \quad (4.6)$$

- Інформаційний критерій Акаіке (ІКА, англ. Akaike information criterion, AIC) – це міра відносної якості статистичних моделей для заданого набору даних. Маючи сукупність моделей для наявних даних, ІКА оцінює якість кожної з моделей відносно кожної з інших моделей. ІКА заснований на теорії інформації: він характеризує відносні оцінки втраченої інформації при застосуванні даної моделі для представлення процесу, що породжує дані. Таким чином, він вказує на компроміс між ступенем узгодженості моделі та її складністю. Припустімо, що ми маємо статистичну модель якихось даних. Нехай L – максимальне значення функції правдоподібності для цієї моделі; і нехай k буде числом оцінюваних параметрів у цій моделі. Тоді значення ІКА для цієї моделі обчислюється так:

$$AIC = 2k - 2\ln(L) \quad (4.7)$$

Для оцінки адекватності моделі також можна використовувати критерій Байєса-Шварца.

4.1.4. Аналіз якості прогнозу

Важливим моментом процесу прогнозування є об’єктивне визначення якості отриманого прогнозу. Оскільки прогнозовані значення – випадкові величини, то для оцінювання їх якості необхідно використовувати декілька статистичних критеріїв. Рис. 4.1 ілюструє часову вісь та відрізки часу, на яких виконується оцінювання моделі і перевірка якості прогнозу.

Рис. 4.1. Види прогнозування за часовим рядом

Наявну вибірку даних доцільно розділити на навчальну та перевірочну. На навчальній вибірці виконується оцінювання параметрів моделі процесу і реалізується так званий „історичний" прогноз, який дає змогу встановити якість однокрокового прогнозу на цьому участку ряду. Прогноз на перевірочній частині вибірки даних в науковій літературі називають ще прогнозом *ex post*. В різних емпіричних дослідженнях рекомендують залишати для перевірки (5 – 40) % значень ряду даних. Хоча при аналізі коротких рядів доцільно значно більшу частину ряду використовувати для оцінювання параметрів моделі. Прогнозування значень поза вибіркою даних називають прогнозом *ex ante* (рис. 4.1).

Як правило, для оцінювання якості прогнозів використовують множину взаємно доповнюючих статистичних критеріїв. Наприклад, значення середньоквадратичної похибки залежить від масштабу даних, а тому недостатньо використовувати тільки цей статистичний параметр для аналізу якості прогнозу. Розглянемо деякі статистичні критерії якості прогнозу та їх призначення.

Обов'язковим етапом прогнозування є точність та обґрунтованість прогнозів. На цьому етапі використовується сукупність критеріїв, підходів і процедур, які можуть оцінити якість прогнозу.

Показниками якості прогнозу є такі:

- Середній квадрат похибок моделі. Для обчислення середнього квадрату похибок (MSE) все окремі залишки регресії зводяться в квадрат,

підсумовуються, а сума ділиться на загальне число похибок:

$$MSE = \frac{\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}{n} \quad (4.8)$$

Квадратний корінь з цієї величини позначається як RMSE (Root Mean Squared Error):

$$RMSE = \sqrt{\sum_{i=1}^{n} \frac{(y_i - \hat{y}_i)^2}{n}} \quad (4.9)$$

- Абсолютна помилка прогнозу може бути визначена як різниця між фактичним значенням (y) і прогнозом (y*):

$$\Delta_{pr} = y_t - y^* \quad (4.10)$$

Середнє абсолютне значення помилки становитиме:

$$\bar{\Delta}_{pr} = \frac{\sum_{t=1}^{n} |y_t - y_t^*|}{n} \quad (4.11)$$

- Коефіцієнт нерівності Тейла. Коефіцієнт нерівності Тейла U – це важливий індикатор якості моделі і прогнозу; за означенням, $0 \le U \le 1$. Якщо $U = 1$, то модель має практично нульові (неприйнятні) прогнозуючі властивості, що випливає з формули для обчислення U:

$$U = \frac{\sqrt{\frac{1}{N}\sum_{k=1}^{N}[y(k) - \hat{y}(k)]^2}}{\sqrt{\frac{1}{N}\sum_{k=1}^{N} y^2(k)} + \sqrt{\frac{1}{N}\sum_{k=1}^{N} \hat{y}^2(k)}} . \quad (4.12)$$

При $U = 0$ прогнозовані значення співпадають з фактичними значеннями ряду – модель ідеальна. Тобто U дає можливість встановити придатність моделі (методу) в принципі для оцінювання прогнозу[14].

– Середня абсолютна похибка в процентах (САПП) – це середнє абсолютних значень похибок оцінок прогнозу в процентах відносно фактичного значення показника:

$$САПП = \frac{1}{N}\sum_{k=1}^{N} \frac{|y(k) - \hat{y}(k)|}{|y(k)|} \times 100\% = \frac{1}{N}\sum_{k=1}^{N} \frac{|e(k)|}{|y(k)|} \times 100\%, \quad (4.13)$$

або у випадку прогнозування на s кроків відносно $k-$го моменту:

$$САПП = \frac{1}{s}\sum_{i=1}^{s}\frac{|y(k+i)-\hat{y}(k+i,k)|}{|y(k+i)|}\times 100\% = \frac{1}{s}\sum_{i=1}^{s}\frac{|e(k+i)|}{|y(k+i)|}\times 100\%. \quad (4.14)$$

Оскільки ця міра характеризує відносну якість прогнозу, то її використовують, в основному, для порівняння точності прогнозів різнорідних об'єктів (процесів) прогнозування. Однак, вона є завжди корисною при виконанні порівняльного аналізу якості прогнозування одного й того ж процесу різними методами, оскільки відносна міра є чіткою і зрозумілою для дослідника і практичного користувача. Типові значення САПП та їх пропонована інтерпретація наведені в таблиці 4.1 [11].

Якщо в формулах (4.13) і (4.14) $y(k)$ або $y(k+i)$ прямують до нуля, то значення САПП прямуватиме до нескінченності. Про це необхідно пам'ятати при застосуванні даного критерію якості прогнозу. Для того щоб виконати обчислення даного критерію в таких випадках, нульові значення $y(k)$ або $y(k+i)$ необхідно пропускати з відповідним корегуванням значення N або s. Можливо, що такий підхід не відповідає деяким вимогам статистичного аналізу даних, але він дає можливість наближено і більш повно виконати аналіз якості прогнозування.

Таблиця 4.1

Інтерпретація типових значень критерію САПП

САПП, %	Інтерпретація
< 10	Висока точність
10 – 20	Хороша точність
20 – 50	Задовільна точність
> 50	Незадовільна (неприйнятна) точність

– Середня похибка – це не відносний показник, вона характеризує ступінь

зміщення прогнозованих значень від фактичних і розраховується за формулою:

$$СП = \frac{1}{N}\sum_{k=1}^{N}[y(k) - \hat{y}(k)] = \frac{1}{N}\sum_{k=1}^{N} e(k), \quad (4.15)$$

або

$$СП = \frac{1}{s}\sum_{i=1}^{s}[y(k+s) - \hat{y}(k+s,k)]. \quad (4.16)$$

Очевидно, що СП буде зменшуватись у випадках, коли похибки мають різні знаки.

Середню похибку в процентах (СПП) обчислюють за виразом:

$$СПП = \frac{1}{N}\sum_{k=1}^{N}\frac{[y(k) - \hat{y}(k)]}{y(k)} \times 100\%, \quad (4.17)$$

або

$$СПП = \frac{1}{s}\sum_{i=1}^{s}\frac{[y(k+s) - \hat{y}(k+s,k)]}{y(k+s)} \times 100\%. \quad (4.18)$$

СПП також характеризує *зміщеність* прогнозу. Якщо втрати при прогнозуванні, зв'язані із завищенням фактичного майбутнього значення, врівноважуються заниженням, то ідеальний прогноз має бути незміщеним. В такому випадку СП і СПП повинні прямувати до нуля. Очевидно, що нуль – це ідеальне значення і забезпечити його на практиці неможливо. Емпірично встановлено, що прийнятними значеннями для СПП (так само як і для САПП) є $\leq 5\%$.

– Максимальна абсолютна похибка. Очевидно, що максимальна абсолютна похибка (МАП) може бути визначена як

$$МАП = \max_{k}\{|\, y(k) - \hat{y}(k)\,|\},\ 1 \leq k \leq N, \quad (4.19)$$

– Мінімальна абсолютна похибка (МіАП) визначається як

$$МіАП = \min_{k}\{|\, y(k) - \hat{y}(k)\,|\},\ 1 \leq k \leq N, \quad (4.20)$$

Критерії МАП і МіАП також можуть бути корисними при виконанні порівняльного аналізу кількох методів прогнозування, особливо якщо нас цікавлять максимально або мінімально можливі відхилення прогнозів від фактичних значень на заданому інтервалі.

4.2. Комбінування прогнозів, отриманих за різними методами

4.2.1. Усереднення прогнозів (вагові коефіцієнти однакові)

Для двох методів прогнозування середнє визначається досить просто:

$$\hat{y}_c(k) = \frac{\hat{y}_1(k) + \hat{y}_2(k)}{2}, \tag{4.21}$$

де $\hat{y}_c(k)$ – комбінований прогноз; $\hat{y}_1(k)$, $\hat{y}_2(k)$ – прогнози, отримані різними методами. Якщо окремі прогнози не зміщені (це повинен забезпечувати метод прогнозування), то комбінованний прогноз також буде незміщеним. Похибка комбінованого прогнозу:

$$e_c(k) = y(k) - \hat{y}_c(k) = y(k) - \frac{\hat{y}_1(k) + \hat{y}_2(k)}{2} = \frac{e_1(k) + e_2(k)}{2}, \tag{4.22}$$

де $y(k)$ – фактичне значення прогнозованої змінної.

Дисперсія похибки комбінованого прогнозу:

$$\text{var}\left[\frac{e_1(k) + e_2(k)}{2}\right] = E\left[\frac{e_1(k) + e_2(k)}{2}\right]^2 = \frac{1}{4}E\left[e_1^2(k) + 2e_1(k)e_2(k) + e_2^2(k)\right] =$$

$$= \frac{1}{4}\left\{E\left[e_1^2(k)\right] + 2E\left[e_1(k)e_2(k)\right] + E\left[e_2^2(k)\right]\right\} = \tag{4.23}$$

$$= \frac{1}{4}\left[\sigma_1^2 + 2\frac{E[e_1(k)e_2(k)]}{\sigma_1\sigma_2}\sigma_1\sigma_2 + \sigma_2^2\right] = \frac{\sigma_1^2 + 2\rho\sigma_1\sigma_2 + \sigma_2^2}{4},$$

Таким чином, дисперсія комбінованого прогнозу обчислюється за виразом:

$$\sigma_c^2 = \frac{\sigma_1^2 + \sigma_2^2 + 2\rho\sigma_1\sigma_2}{4}, \tag{4.24}$$

де ρ – коефіцієнт кореляції між похибками прогнозу. Якщо похибки прогнозування за двома моделями незалежні, тобто, $\rho = 0$, то формула (4.24) спрощується:

$$\sigma_c^2 = \frac{\sigma_1^2 + \sigma_2^2}{4}, \tag{4.25}$$

Таким чином, якщо дисперсії близькі за значеннями і незалежні, то

дисперсія комбінованої похибки буде значно меншою будь-якої з двох дисперсій. Наприклад, нехай $\sigma_1^2 = \sigma_2^2 = 100$:

$$\sigma_c^2 = \frac{100 + 100}{4} = 50.$$

Але навіть при існуванні досить високої кореляції між похибками прогнозування дисперсія похибки комбінованого прогнозу буде меншою ніж дисперсія кожного методу окремо. Наприклад, нехай $\sigma_1^2 = \sigma_2^2 = 100$ і $\rho = 0.8$:

$$\sigma_c^2 = \frac{\sigma_1^2 + \sigma_2^2 + 2\rho\sigma_1\sigma_2}{4} = \frac{100 + 100 + 2 \cdot 0.8 \cdot 10 \cdot 10}{4} = 90.$$

Навіть в цій ситуації спостерігається зменшення дисперсії похибки прогнозу після усереднення оцінок, отриманих за двома методами.

Однак, ситуація змінюється у випадку, коли дисперсії індивідуальних похибок сильно відрізняються. Наприклад, нехай $\sigma_1^2 = 100,\ \sigma_2^2 = 16$ і $\rho = 0.8$:

$$\sigma_c^2 = \frac{\sigma_1^2 + \sigma_2^2 + 2\rho\sigma_1\sigma_2}{4} = \frac{100 + 16 + 2 \cdot 0.8 \cdot 10 \cdot 4}{4} = 45.$$

Таким чином, якщо дисперсії похибок сильно відрізняються, то просте усереднення результатів не потрібно робити,. Висновок такий: просте усереднення можна застосовувати у випадках, коли дисперсії індивідуальних похибок прогнозування приблизно рівні або не дуже сильно відрізняються за своїми значеннями.

4.2.2 Зважене усереднення прогнозів

Якщо інформація щодо характеристик індивідуальних прогнозів відсутня, то можна присвоїти різні вагові коефіцієнти окремим прогнозам на основі суб'єктивних або експертних суджень:

$$\hat{y}_c(k) = w_1\, \hat{y}_1(k) + w_2\, \hat{y}_2(k), \qquad (4.26)$$

де w_1, w_2 – вагові коефіцієнти. Очевидно, що більші значення вагових коефіцієнтів необхідно присвоювати тим індивідуальним прогнозам, які мають меншу дисперсію похибок. При цьому для коректності обчислень необхідно,

щоб виконувалась умова: $w_1 + w_2 = 1$.

4.2.3 Вибір вагових коефіцієнтів за допомогою похибок прогнозів

Як правило, похибки прогнозів для конкретних моделей і процесів відомі, або їх можна визначити. Це дає можливість об'єктивно підійти до розв'язку задачі вибору вагових коефіцієнтів. Оскільки моделі, які дають меші суми квадратів похибок прогнозів, генерують якісніші прогнози, то логічно прийняти цю міру за основу для визначення вагових коефіцієнтів. Позначимо суму квадратів похибок прогнозування (для історичного прогнозу) через

$$sse = \sum_{k=1}^{N} e^2(k), \tag{4.27}$$

Тепер можна записати вирази для вагових коефіцієнтів окремих прогнозов:

$$w_1 = \frac{1/sse_1}{1/sse_1 + 1/sse_2}, \tag{4.28}$$

$$w_2 = \frac{1/sse_2}{1/sse_1 + 1/sse_2}, \tag{4.29}$$

де sse_1, sse_2 – суми квадратів похибок для кожного з методів, що використовуються в даному випадку.

Наприклад, нехай $sse_1 = 100,\ sse_2 = 16$:

$$w_1 = \frac{1/100}{1/100 + 1/16} = \frac{0.01}{0.01 + 0.0625} = 0.1379,$$

$$w_2 = \frac{1/16}{1/100 + 1/16} = \frac{0.0625}{0.01 + 0.0625} = 0.8621.$$

Таким чином, ми (об'єктивно) присвоїли більший ваговий коефіцієнт точнішому методу прогнозування. При цьому $\sum w_i = 1$, що необхідно для досягнення коректності застосування методу.

4.3 Розробка концепції інформаційної технології для моделювання та прогнозування фінансових

В сучасному світі галузь інформаційних технологій розвивається швидкими темпами, при цьому кожний рік вони зростають. В останні роки активно розвиваються сервісні ресурси, якими можна користуватись безпосередньо через мережу Інтернет. Можливість використання ресурсу у мережі Інтернет стало одним із основних факторів, що відокремлює інформаційні ресурси, які є сучасними та мають більші можливості подальшого розвитку, від програмного забезпечення (ПЗ), що має меншу практичну значущість і подальший розвиток якого потребує більших зусиль.

Більшість програмного забезпечення для виконання байєсівського аналізу даних, описаного в другому та третьому розділах, потребує певних знань та навичок від користувача, а деколи і навичок програмування. Майже відсутнє ПЗ для застосування фільтрації даних, заповнення пропусків даних, комбінування прогнозів, використання якого можливе у мережі «Інтернет». Після виконання докладного огляду інформаційних систем не виявлено жодного ресурсу, що свідчить перший розділ, який забезпечив би можливість прогнозування значень кредитних та ринкових ризиків у реальному часі.

Основними задачами розробки інформаційної системи є:

- реалізація описаних у попередніх розділах алгоритмів оцінювання та прогнозування фінансово-економічних процесів;
- забезпечення можливості доступу користувачів до системи через мережу Інтернет;
- створення можливості використання іншими розробниками сервісу прогнозування завдяки використанню веб–сервісів;
- розробка «простого» та інтуїтивно зрозумілого інтерфейсу користувача, створеного за рахунок автоматизованих та прихованих від користувача опцій налаштувань;
- забезпечення можливості подальшого легкого розширення та

модифікації функцій системи.

Для цього постає задача вибору інструментів для розробки та технологій для реалізації інформаційної системи, яка б задовольняла поставленим вимогам. Сьогодні однією з найпотужніших мов програмування є JavaScript, Python, оскільки використання розроблених на цій мові систем можливе на будь-якій операційній системі, а кожен з розробників може збільшити можливості мови за рахунок використання або створення власних бібліотек та їх розміщення у мережі. В цьому підрозділі докладно описаний інструментарій, який застосовується у розробленій системі. Розглянуто етапи проектування системи та її програмної реалізації. Наведені всі модулі, частина користувача, серверна частина та база даних (БД).

4.3.1 Концепція інформаційної технології

Перед початком проектування інформаційної технології необхідно розглянути загальний принцип роботи системи. Доступ до системи здійснюється через мережу Інтернет. Система реалізована на основі клієнт-серверної архітектури. Користувачами можуть бути як користувачі графічного інтерфейсу системи, так і зовнішні системи. Обробку запитів, що надходять до системи, здійснює веб-сервер системи. Веб-сервер функціонує наступним чином: приймає запити від користувача; формує та відправляє запити на проведення розрахунків до сервера застосувань; отримує від сервера застосувань результати розрахунків; формує відповідь користувачеві (рис. 4.1).

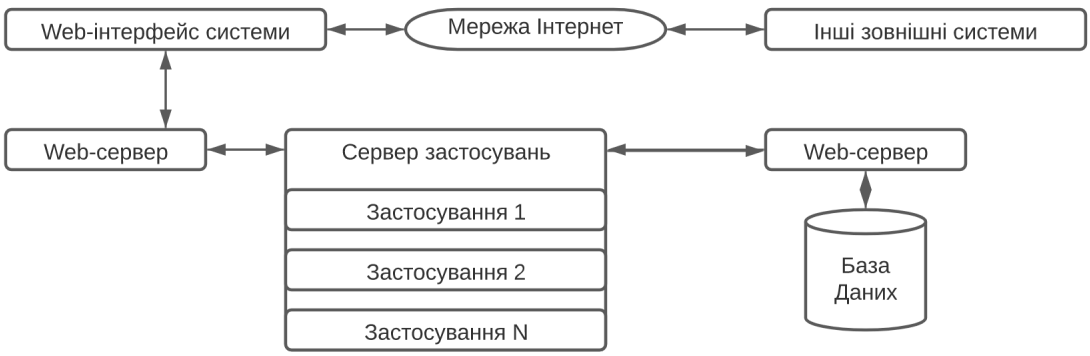

Рис. 4.1. Структура інформаційної системи

4.3.2 Проектування інформаційної технології

Проектування системи починається з виділення сутностей предметної області та їх взаємозв’язків. Після цього визначається прийнятний інструментарій і технології, оскільки вони впливають на етап проектування. На практиці у процесі створення сучасної інформаційної системи необхідно реалізувати такі артефакти:

- архітектуру системи;
- діаграму послідовності виконання операцій (Sequence diagram);
- модель сценаріїв користувача (User story);
- модель прецедентів (сценарій використання – Use case);
- бізнес-процеси;
- модель «сутність - зв'язок» (entity-relationship diagram);
- модель відповіді системи на дію користувача (Acceptance Scenario);
- графічне представлення інтерфейсу користувача (Graphical user interface).

4.3.3 Архітектура інформаційної технології

Існують різноманітні підходи архітектурного проектування. Можна виділити декілька етапів, спільних для всіх процесів архітектурного проектування:

1. *Структурування системи.* Програмна система структурується у вигляді множини відносно незалежних підсистем.

2. *Моделювання керування.* Розроблюється базова модель керування взаємовідносинами між частинами системи.

3. *Модульна декомпозиції.* Кожна визначена на першому етапі підсистема розбивається на окремі модулі.

Як правило, ці етапи змішуються та накладаються один на одного. При проектуванні інформаційної системи бажано розробити модульно-структурну архітектуру системи, що планується реалізувати. Технічна архітектура буде представлена нижче після вибору інструментарію реалізації. Розглянемо

модульно-структурну архітектуру.

Перше, що потрібно враховувати, розроблюючи сучасне архітектурне рішення, – це забезпечити можливість взаємодії системи із зовнішніми ресурсами та користувачами. Для того щоб забезпечити якомога гнучкіше рішення, реалізована система повинна автоматично взаємодіяти із зовнішніми ресурсами через інструментарій веб-сервісів та мати власний інтерфейс користувача.

Описані в попередніх розділах моделі для прогнозування фінансових ризиків можуть бути застосовані для опису різних фінансових індексів. Значення фінансових показників, наприклад курсу валют, оновлюються кожний день і їх можна отримати у форматі CSV або JSON документу. Для коректної реалізації процедури оцінювання необхідна вибірка з декількох сотень значень, тому значення курсу валют, що надходять від сторонніх ресурсів, необхідно зберігати у базі даних, що значно зменшить навантаження на систему. Функціонал побудови графіків також потребує збережених даних. Крім того, значення, що надаються у форматі CSV, потребують розбору (або так званого парсингу, що дозволить їх структурувати).

При використанні системи через наявний інтерфейс необхідно надати користувачу можливість перегляду поточної інформації про фінансові показники. Дані мають бути представлені у вигляді таблиці. Для кращого сприйняття числової інформації повинна бути можливість її представлення у вигляді графіків.

Винесена в окремий модуль процедура прогнозування не є ресурсномісткою і повинна запускатись при кожному зверненні користувача. Даний функціонал повинен надаватись користувачу графічного інтерфейсу системи, а також іншим зовнішнім системам. Для того щоб не дублювати код цього функціоналу, необхідно приділити увагу розміщенню такого модуля з технічної точки зору, що дозволить створити незалежні модулії, які будуть передавати результати прогнозування відповідно користувачу інтерфейсу системи або зовнішній системі.

На рисунку 4.2 представлена розроблена модульно-структурна архітектура ІС.

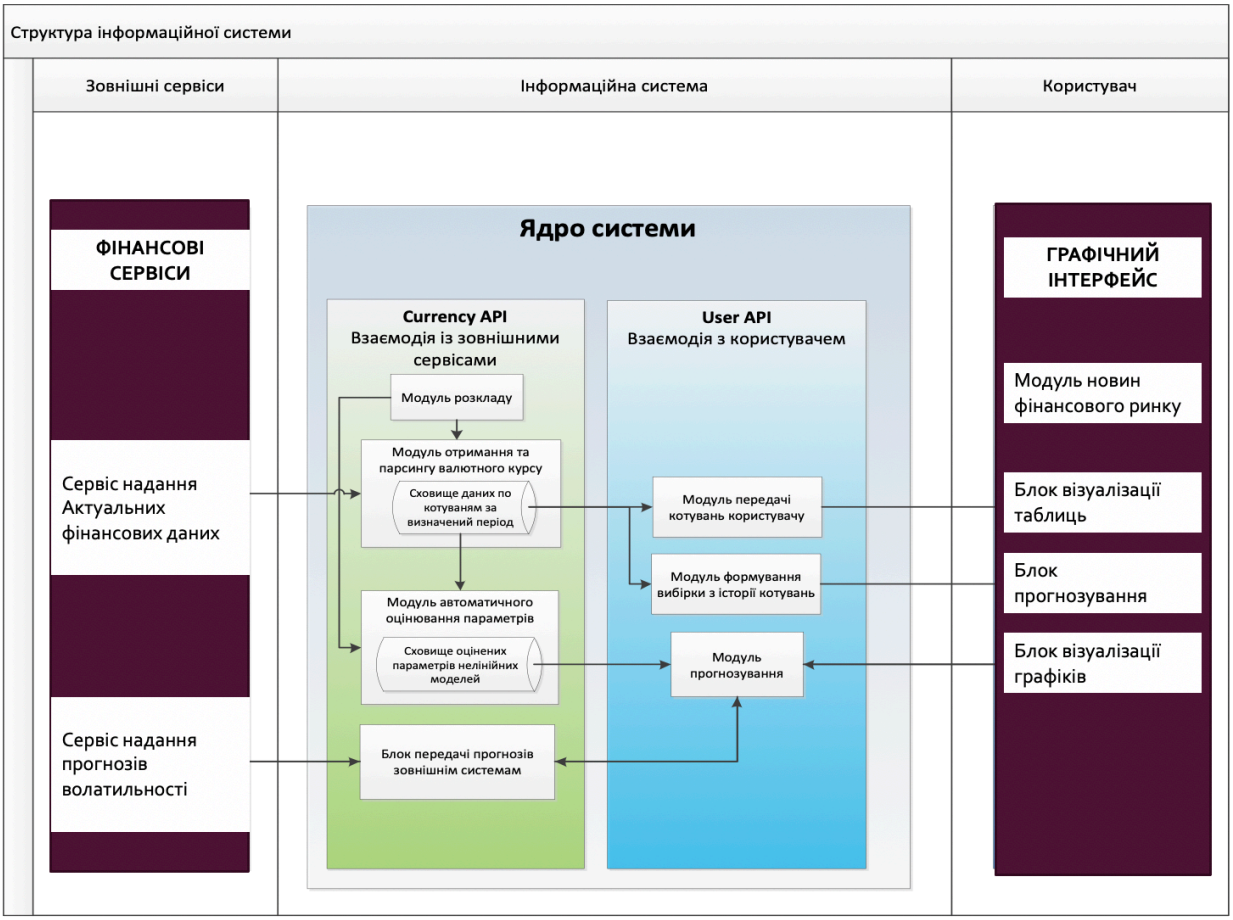

Рис. 4.2. Модульно-структурна архітектура ІС

4.3.4 Діаграма послідовності взаємодії об’єктів технології

Діаграма послідовності взаємодії об’єктів відіграє важливу роль при проектуванні системи. Вона надає можливість краще зрозуміти послідовність взаємодії користувача з системою та послідовність взаємодії об’єктів системи.

Виконання процедури прогнозування потребує оцінок параметрів відповідних моделей. В свою чергу, для оцінювання параметрів бажано мати останні значення фінансових показників. Значення валютного курсу оновлюється за деякий час до початку нової доби. Першим кроком є отримання валютного курсу із зовнішніх веб-сервісів. Значення, що надійшли, необхідно записати до БД. Процедура оцінювання параметрів може проходити і без останніх значень валютного курсу, для цього необхідна вибірка за деякий проміжок часу. З метою кращого врахування останніх змін на фінансовому ринку

бажано проводити оцінювання параметрів після надходження значень фінансових показників.

На рис. 4.3 зображена діаграма послідовності взаємодії між об’єктами системи.

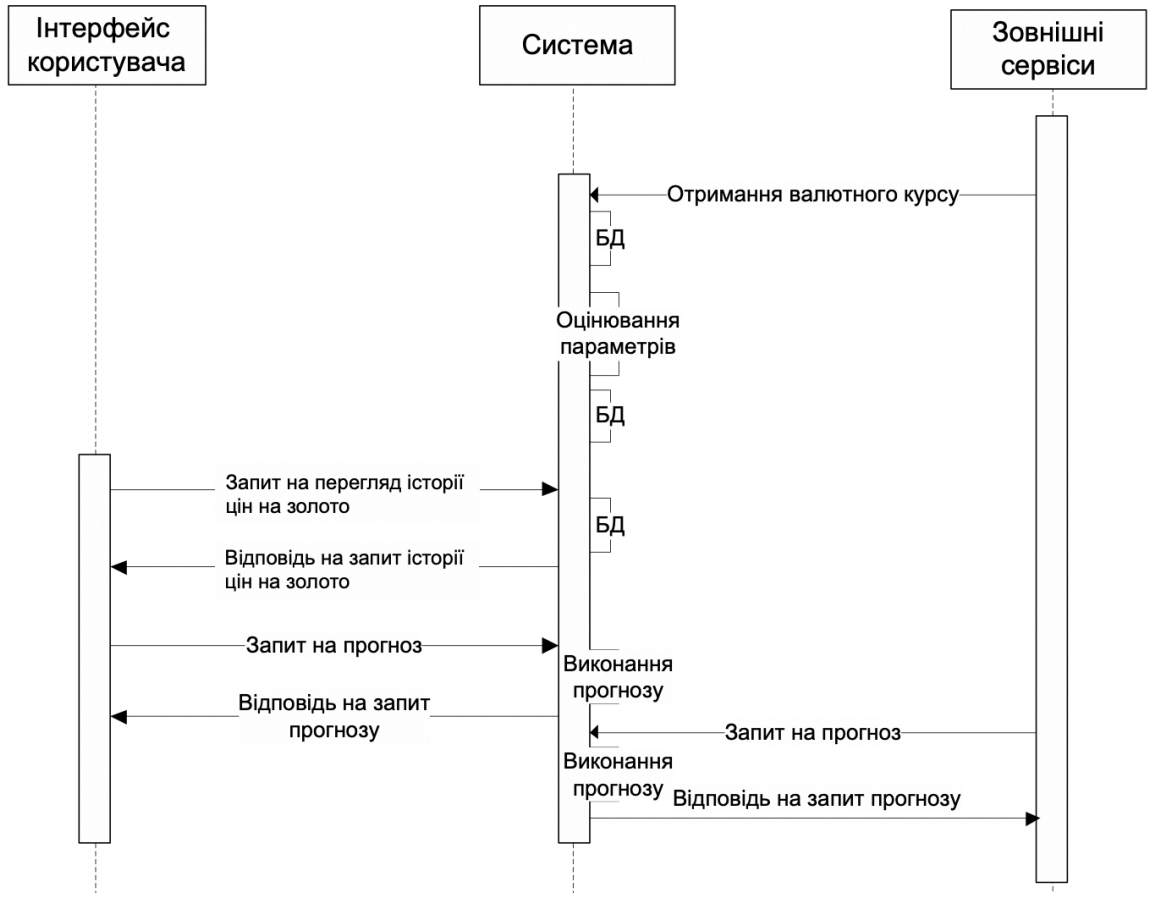

Рис. 4.3. Діаграма послідовності взаємодії між об’єктами системи

4.3.5. Реалізація інформаційної технології

Важливим етапом, після проектування системи є її реалізація. Спроектована система може бути реалізована на основі різного інструментарію.

4.3.5.1 Інструментарій інформаційної технології

Перед початком реалізації системи необхідно визначитись з інструментарієм розробки. Для реалізації спроектованої системи необхідно визначитись із такими елементами:

- множина засобів розробки;

- сервер застосувань;
- база даних.

Крім зазначених елементів, при розробці сучасних промислових рішень необхідно використовувати інструмент для управління проектами та засоби тестування. Невелика кількість мов програмування задовольняє вимогам, що висуваються до сучасних інформаційних систем. Серед них мова програмування JavaScript для реалізації частини користувача та Python для реалізації серверної частини. Зокрема, реалізовані на JavaScript та Python рішення можуть виконуватись у будь-якій операційній системі. Важливо також і те, що існує можливість розширення можливостей мови сторонніми розробниками шляхом написання бібліотек та розміщення їх у мережі Інтернет. Набір засобів розробки на мові JavaScript надає комплект фреймворків (React, Angular). Для мови програмування Python - набір бібліотек з відкритим кодом.

Для реалізації спроектованого функціоналу необхідно застосувати серверну архітектуру. Використання фреймворків та бібліотек в основному, у високопродуктивних проектах, надає системі надійності, масштабованості та гнучкості. Вона орієнтована на використання через Інтернет.

Контейнери надають базові сервіси (технології) для розвернутих на їх основі компонентів. Це надає можливість розробнику концентруватись на бізнес-логіці системи, а не на вирішенні технічних проблем.

Реалізовані за даною специфікацією системи, розмішуються на серверах застосувань. Системою керування БД обрано MySQL. Даний сервер БД характеризується стійкістю та простотою використання.

Спроектована ІС має розгорнуту архітектуру та використовує велику кількість бібліотек. Це потребує використання інструментарію для управління проектами. При розробці великих корпоративних рішень процес виявлення некоректної роботи системи ускладнюється. На допомогу приходить інструментарій тестування. Значного поширення набула бібліотека для тестування програмного забезпечення JUnit.

Таким чином, система реалізована з залученням такого технологій:

- Java Script (React, Angular);
- Python;
- MySQL 5.5;

4.3.5.2 Технічна архітектура інформаційної технології

Після створення структурно-модульної архітектури та визначення технологій розробки можна переходити до реалізації технічної архітектури ІС. Вона допомагає розробникам краще зрозуміти, які фізичні елементи необхідно реалізувати та як вони повинні взаємодіяти.

Завдяки високому сучасному рівню розвитку інформаційних технологій існує велика кількість можливих варіантів реалізації системи. За допомогою мережі Інтернет система може знаходитись фізично в різних куточках світу і функціонувати як єдина система. Це досягається шляхом розділення функціоналу системи для їх роботи на різних серверах. З метою надійної роботи весь функціонал системи розміщено на одному сервері. БД може потребувати дій, пов’язаних із створенням резервної копії. Має сенс розмістити БД на окремому сервері.

4.4 Аналіз результатів моделювання кредитних ризиків

Складність моделювання кредитного ризику обумовлена тим, що задача оцінювання кредитоспроможності відноситься до типу слабкоструктурованих проблем, основними рисами яких є об'єктивна наявність у їх складі як якісних, так і кількісних показників. Крім того, значна частина банківської інформації не піддається точному й формалізованому опису, але може бути представлена у формі нечіткого подання.

Пропонується процес аналізу моделювання кредитного ризику банку та використання регресійних, нейронечітких моделей, дерев рішень та мереж Байєса для управління кредитним ризиком. Ефективність цього підходу особливо виявляється тоді, коли процеси, що розглядаються, досить складні для аналізу за допомогою класичних кількісних методів, або коли джерела

інформації інтерпретуються якісно, неточно або невизначено, тобто коли процеси є слабко структурованими.

Процес аналізу представляє собою декілька етапів, що зображені на рисунку 4.4.

Рис. 4.4 Процес аналізу статистичних даних

Експеримент 1.

Етап 1. База знань даних, опис даних для експерименту.

Для початкового аналізу та побудови математичних моделей для оцінювання кредитоспроможності було обрано вибірку із 3350 клієнтами, які характеризувалися такими атрибутами: *x1*– стать позичальника, *x2*–вік

позичальника, *x3* - сімейний статус, *x4*- кількість дітей, *x5* - наявність особистого житла, *x6* - працює / не працює, *x7*–освіта, *x8* - вид роботи, *x9*–наявність гарантованої особи,*x10* - сума кредиту. Вихідну змінну, оцінку рівня кредитного ризику, позначимо через *y* (результат аналізу).

Етап 2. Заповнення пропусків даних.

Інформація, що збирається в результаті вибіркових даних, потребує обробки з метою розповсюдження на генеральну сукупність. На підставі консолідованого масиву всіх даних, отриманих у ході обстеження, визначаються значення узагальнюючих показників.

Найпростішим рішенням оброблення неповних даних є виключення некомплектних спостережень, що містять пропуски, і подальший аналіз отриманих таким способом "повних" даних. Зрозуміло, що такий підхід призводить до сильної відмінності висновків, зроблених за наявності в даних пропусків і за їх відсутності. Тому перспективнішим є інший шлях – заповнення пропусків перед аналізом масиву даних. Такий підхід має такі переваги: чітке представлення структури даних; обчислення необхідних пропущених значень; впевнена інтерпретація результатів аналізу, оскільки можна спиратися на традиційні характеристики та сумарні значення. Для початкових даних аналізу кредитоспроможності були додані штучні пропуски у співвідношенні 5%, 10%, 15% та 20%.

Заповнення пропусків було виконано за допомогою моделі регресії, середнім значенням за вибіркою, середнім значенням між першим попереднім та першим наступним значенням, середніми п’ятьма попередніми значеннями, що розраховані з використанням вагових коефіцієнтів за принципом експоненційного згладжування.

У таблиці 4.2 наведено значення похибки MAPE при заповненні пропусків даних кредитоспроможності різними методами.

Таблиця 4.2

Значення похибок MAPE при заповненні пропусків

Метод заповнення пропусків	Часовий ряд	Пропущено значень, %			
		5	10	15	20
Модель регресії	Заробітна плата	16,12	17,16	16,69	18,55
	Сума кредиту	4,92	5,01	5,26	4,99
Середнє за вибіркою	Заробітна плата	22,15	24,19	24,78	26,07
	Сума кредиту	28,35	27,87	31,12	32,56
Середнє між першим попереднім та першим наступним значеннями	Заробітна плата	10,14	11,43	10,45	12,07
	Сума кредиту	4,43	4,78	5,13	5,56
Середнє п’яти попередніх значень, розрахованих з використанням вагових коефіцієнтів за принципом експоненційного згладжування	Заробітна плата	17,12	18,48	19,34	21,22
	Сума кредиту	7,96	8,24	8,87	9,11

У експерименті, на основі таблиці 4.1, використані часові ряди із штучними пропусками даних у кількості від 5% до 20% від загальної кількості значень часового ряду. Як показано у таблиці 4.1, значення похибки MAPE суттєво відрізняється для досліджуваних часових рядів. Однак, найбільш ефективними виявилися методи заповнення пропусків значеннями, спрогнозованими із використанням моделі регресії (MAPE = 4,92%), та значеннями, розрахованими як середнє між першим попереднім та першим наступним значеннями (MAPE = 4,43%).

Етап 3. Оцінювання інформативності даних та структурування вибірки

При наданні кредитів банк зобов’язаний дотримуватись основних принципів кредитування, зокрема проводити аналіз кредитоспроможності

позичальників, дотримуватись встановлених банком вимог щодо концентрації ризиків [3, 5].

Найпершим і найбільш значущим моментом при видачі кредиту є оцінка кредитоспроможності. І якщо гарантом платоспроможності фізичної особи може бути високий рівень заробітної плати, зарплатня поручителів або застава власного майна , то для юридичних осіб гарантія такого роду існує не завжди повною мірою. Також існує кілька обхідних варіантів, а суми позик при цьому можуть бути значно вищі.

Різні фактори можуть впливати на кредитоспроможність позичальника:

фінансове становище – рівень забезпеченості власними коштами, ліквідні активи, фінансова дисциплінованість, склад родини. Для аналізу та знаходженню взаємозв'язків між атрибутами даних використовується ряд статистик для оцінювання значущості (рис. 4.5):

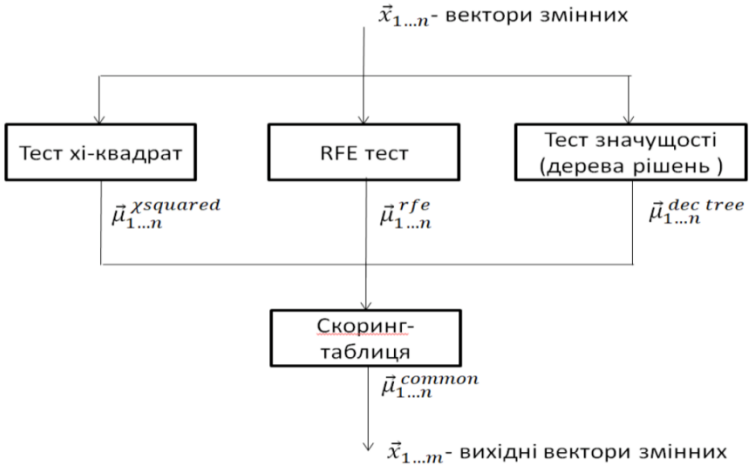

Рис. 4.5 Схема встановлення зв'язків між атрибутами

Першим методом оцінки інформативності моделі є загальний статистичний тест. Даний підхід полягає у встановленні залежності між екзогенною змінною і цільовою змінними. Знаходження зв'язку між змінними проводиться за допомогою статистики хі-квадрат. Розрахувавши очікувану величину, ми можемо визначити залежність між двома величинами, та зробити порівняння з іншими тестами на оцінювання інформативності даних.

В якості наступних підходів розглядається метод RFE – рекурсивне

виключення елементів та оцінювання значущості на основі дерев рішень, що імплементує оцінювання значущості екзогенних змінних на основі дерев прийняття рішень. Результати експериментів та аналіз показників наведено у таблиці 4.3.

На цьому етапі вибираються значущі змінні для їх введення у модель. Отже, відбір змінних для прогнозування відбувається в 3 кроки.

На першому кроці відбувається вибір за загальним статистичним тестом Хі-квадрат. При задані порогу в 20, було обрано 6 змінних:x4- кількість дітей, x5 - наявність особистого житла, x6 - працює / не працює, x7–освіта, x8 - вид роботи, x10 - сума кредиту. Кожна із вище зазначених змінних має по одному балу для загального розрахунку значущості.

На другому кроці відбувається вибір за методом рекурсивного виключення змінних. Було обрано дефолтне значення кількості змінних – 6.

Таким чином, обрано такі змінні: x1– стать позичальника, x2–вік позичальника, x5 - наявність особистого житла, x6 - працює / не працює, x8 - вид роботи,x10 - сума кредиту. Цим змінним присвоєно по 1 балу за відбір до моделі за методом RFE.

Третій крок представляє визначення значущості змінних за допомогоюдерев рішень. До аналізу на основі дерев рішень включено такі результати: x2–вік позичальника, x3–сімейний статус, x4–кількість дітей,x5–наявність особистого житла, x6–працює / не працює, x7–освіта, x8– вид роботи,x10– сума кредиту. Цим змінним було присвоєно один бал за версією третього методу відбору, всі іншим змінним присвоєно 0.

Базуючись на отриманих розрахунках за статистичними методами, необхідно розрахувати фінальне значення для кожної зі змінних, і на основі загальної функції активації вибору отримаємо, чи варто обирати дану зміну до фінальної моделі. Отже, розраховуємо загальний бал за композицією трьох методів відбору. В результаті отримано значення, подані в таблиці 4.3:

Таблиця 4.3

Розрахунок загального скорингового балу для кожного атрибута

Атрибут	Загальний статистичний тест	Метод рекурсивного виключення змінних	Тест значущості на основі дерев рішень	Загальний бал
x1 – стать позичальника	0	1	0	1
x2–вік позичальника	0	1	1	2
x3 - сімейний статус	0	0	1	1
x4- кількість дітей	1	0	1	2
x5 - наявність особистого житла	1	1	1	3
x6 - працює / не працює	1	1	1	3
x7–освіта	1	0	1	2
x8 - вид роботи	1	1	1	3
x9–наявність гарантованої особи	0	0	0	0
x10 - сума кредиту	1	1	1	3

Обрано дефолтне значення фінальної функції активації «1», тобто змінні із загальним балом в 2 і більше вважаються значущими і мають бути введені в модель:x2– вік позичальника, x4- кількість дітей,x5 - наявність особистого житла, x6 - працює / не працює, x7– освіта, x8 - вид роботи,x10 - сума кредиту. Всі інші змінні буде відкинуто [2].

Етап 4. На основі отриманих характеристик при інформативному оцінюванні змінних було побудовано моделі та наведено їх точність у таблиці 4.4. В якості моделей було обрано логіт модель, системи на основі нечітких правил з трикутною, трапецевидною, гаусівською, сігмоідною та дзвіноподібною функціями належности, нейрона мережа зворотнього розповсюдження помилки та дерева рішень.

Таблиця 4.4

Показники якості моделі для загальної та скороченої кількості атрибутів

Тип моделі		Точність (загальна кількість)	Точність (скорочена кількість)
Logit model		0,89	0,897
FRBS	TRIANGLE	0,851	0,852
	TRAPEZOID	0,913	0,917
	GAUSSIAN	0,916	0,918
	SIGMOID	0,907	0,91
	BELL	0,932	0,940
BP	10 hidden layers	0,87	0,86
	3 hidden layers	0,958	0,961
Decision Tree - Regression		0,89	0,9

Як видно з таблиці, найкращий результат (точність моделі) було отримано при скороченій кількості атрибутів з використанням системи на основі нечітких правил із дзвоноподібною функцією належності (0,940) та з використанням нейронної мережі зворотного розповсюдження помилки з трьома прихованими шарами (0,961).

Етап 5. Виходячи з того, що скорочена кількість атрибутів показала більш високу точність моделі, прогноз та аналіз результатів виконується з використанням 7 вхідних атрибутів. Для прогнозування використовувались вибірки навчальна та перевірочна у співвідношенні (%): 70/30, 80/20, 90/10.

Таблиця 4.5

Показники прогнозу для скороченої кількості обраних атрибутів

Співвідношення навчальної та перевірочної вибірок, %		70/30	80/20	90/10
Modeltype		Значення MSE		
Logitmodel		0,76658	0,76838	0,7612
FRBS	TRIANGLE	0,6104	0,61050	0,6191
	TRAPEZOID	0,23062	0,23055	0,2303
	GAUSSIAN	0,22044	0,21703	0,2241
	SIGMOID	0,33188	0,33161	0,3297
	BELL	0,1969	0,19156	0,2012
BP	10 hidden layers	0.74196	0.23192	0.7559
	3 hidden layers	0.17475	0.18488	0.2420
Decision Tree - Regression		0.74426	0.75838	0.7497
Logit model (AUC)		0,63068	0,69102	0,6465

Найкращий результат MSE було отримано для систем на основі нечітких правил з дзвоноподібною функцією належності при співвідношенні навчальної та перевірочної вибірки 80/20, що дорівнює 0,19156 та для нейронної мережі зворотного розповсюдження похибки з трьома прихованими шарами при співвідношенні навчальної та перевірочної вибірки 70/30, що дорівнює 0,17475.

Експеримент 2.

Для початкового аналізу та побудови математичних моделей оцінювання кредитоспроможності було обрано вибірку із клієнтами, які характеризувалися такими атрибутами: вік клієнта від 18 до 65 років; стаж роботи від 0 до 49 років; заробітна плата від 0 до 30000 грн.; наявність майна: відсутня, машина, квартира, дім; два та більше з перелічених значень.

Кредитна історія: 1 – є кредити, оплата по яких прострочена; 2 – не було; 3

– є кредити у інших банках та погашаються своєчасно; 4 – є кредити у даному банку та погашаються своєчасно; 5 – немає відкритих кредитів, попередні сплачувались своєчасно; Сума кредиту: 1000 – 100000 (грн.); строк кредитування: від 4 місяців до 5 років. Вихідну змінну, оцінку рівня кредитного ризику, позначимо через *y* (остаточний результат).

Застосування методів моделювання кредитних ризиків

При побудові моделей кредитного ризику використовується вибірка даних про клієнтів, яка поділяється на дві частини – навчальна та перевірочна. В якості цільової змінної буде використовуватися значення результату повернення або неповернення клієнтом кредиту. У разі повернення кредиту значення буде дорівнювати 1, в якості неповернення кредиту значення буде дорівнювати 0. Початкова вибірка складається із записів про 1500 клієнтів з їх характеристиками (характеристики описані вище). Навчальна вибірка буде містити 1200 записів, а перевірочна буде мати інформацію про 300 клієнтів. Вхідними даними будуть такі характеристики: сума кредиту, вік позичальника, дохід позичальника, платіж по кредиту, строк кредиту та освіта позичальника. Характеристики побудованих логіт і пробіт моделей подані в таблиці 4.6.

Таблиця 4.6

Статистичні характеристики якості прогнозів

	Середня квадратична похибка	Середня абсолютна похибка у %	К-т Тейла
Логіт	0.407	13.24	0.494
Пробіт	0.406	13.23	0.493

Отримані статистичні характеристики прогнозування для обох типів моделей дуже близькі між собою. Тобто у даному випадку тип розподілу не має суттєвого значення для побудови моделі та отриманого результату.

Дерева рішень для оцінювання кредитоспроможності позичальників. Дерева рішень зачасту використовують для задач класифікації. Коли до банку звертається клієнт для того, щоб отримати кредит, то за його характеристиками

відбувається класифікація до високого або низького рівня ризику. Завдяки застосуванню дерев рішень, визначаються правила для виконання класифікації основі даних, які зберігаються у навчальній вибірки. Якщо порівнювати дерева рішень з іншими алгоритмами, то перші мають переваги у швидкості та зрозумілості реалізації, навчанні та точності. Для вибірки з 1600 записами було побудовано дерево рішень, яке дозволяє класифікувати нових клієнтів як таких, що повернуть кредит, або не повернуть. Результати побудови дерев рішень зведено для порівняння в таблицю 4.7 та подано у вигляді діаграми (рис. 4.6).

Таблиця 4.7

Застосування дерев рішень для оцінювання кредитоспроможності позичальників

Тип дерева рішень	Загальна точність	Похибки класифікації		
		I-го роду	II-го роду	Сумарна похибка
Chaid	0,713	3,733	11,933	15,666
Exhaustive Chaid	0,705	3,733	13,733	17,466
CRT	0,712	4,4	3,4	7,8
Quest	0,705	5,333	6,1333	11,466

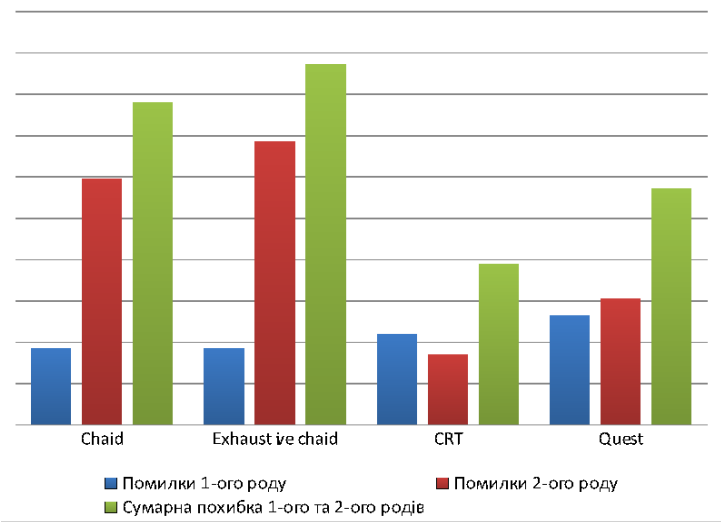

Рис. 4.6 - Помилки 1го та 2го роду

Використання Байєсівських мереж для прогнозування кредитоспроможності індивідуального позичальника

Розглянемо приклад із 1600 записами по кредитним історіям. Запис кожного клієнта характеризується 18 показниками.

За наявними даними було побудовано мережу, яка встановлює зв'язок між характеристиками клієнта і вершиною – подією повернення кредиту. Також визначається ймовірність повернення кредиту новим клієнтом, які прийшли до банку. Так як числові характеристики (сума кредиту, вік, дохід і т. ін.) набувають великої кількості значень (є неперервними), для розв'язання даної задачі використовується гібридна мережа Байєса. За описаною у розділі 2 методикою виконується аналіз проблеми і робиться формалізована постановку задачі. Розглянемо детальніше схему застосування запропонованої методики. Дані, що відносяться до задачі – це статистичні дані за 1600 виданими кредитами, з яких 450 випадків дефолтів, 1150 – повернутих кредитів. Базуючись на основних кроках, отримаємо побудовану за початковими даними структуру мережі Байєса. Вона наочно демонструє зв’язки між даними.

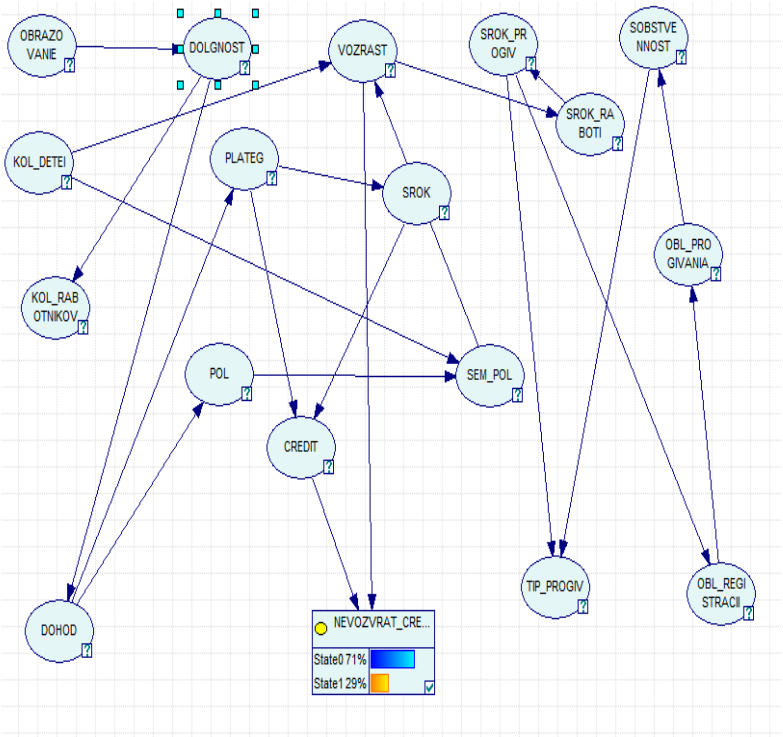

Рис. 4.7 Структура побудованої мережі Байєса

Значення вершин використовується для знаходження порогу відсікання – компромісу між чутливістю та специфічністю моделі. Критеріями вибору порогу відсікання можуть бути:

1) вимога мінімальної величини чутливості (специфічності) моделі;

2) вимога максимальної сумарної чутливості і специфічності моделі, тобто

$$cut-off = \max_k (Se_k + Sp_k) \tag{4.30}$$

3) вимога балансу між чутливістю і специфічністю [6], тобто коли

$$Sp \approx Se: \ cut-off = \min_k | Se_k - Sp_k | \tag{4.31}$$

Таблиця 4.8

Загальна точність моделі та помилки I-го і II-го роду. Розглядаються різні рівні відсікання, які отримані для мережі Байєса

	Повернення кредиту (0)	Дефолт (1)
Cut-off=0,1		
Повернення кредиту (0)	3	237
Дефолт (1)	0	60
Загальна точність моделі = 50,6 %		
Cut-off=0,15		
Повернення кредиту (0)	3	237
Дефолт (1)	0	60
Загальна точність моделі = 50,6 %		
Cut-off=0,2		
Повернення кредиту (0)	22	218
Дефолт (1)	0	60
Загальна точність моделі = 54,6 %		
Cut-off=0,25		
Повернення кредиту (0)	23	217

Дефолт (1)	0	60
Загальна точність моделі = 54,8 %		
Cut-off=0,3		
Повернення кредиту (0)	146	94
Дефолт (1)	3	57
Загальна точність моделі = 77,9 %		

Найбільшої точності моделі було досягнуто при рівні 77 % при встановленні порогу 0,3; при цьому було пропущено 3 значення дефолти та відкинуто 39 % добросовісних позичальників. Значення площі під кривою становить: $AUC = 0{,}879$, а індекс GINI становить: $GINI = 2 * AUC - 1 = 0{,}758$.

Оцінювання кредитоспроможності позичальника з використанням нечіткої логіки

При виконанні аналітичних досліджень визначено перелік факторів, що впливають на формування рівня кредитоспроможності клієнта. Нижче подано перелік вхідних факторів, використаних при розробці системи, зокрема: 1) вік позичальника (x_1); 2) стаж роботи (x_2); 3) місячна заробітна плата (x_3); 4) наявність майна (x_4); 5) наявність кредитної історії (x_5); 6) сума кредиту (x_6); 7) термін кредитування (x_7).

Для оцінювання кредитоспроможності, фактори, що впливають на формування рівня кредитоспроможності клієнта можуть приймати такі значення: вік 18 до 65 років; стаж роботи від 0 до 49 років; заробітна плата від 0 до 30000 грн.; наявність майна: відсутня; машина; квартира; дім; два та більше з перелічених значень. Наявність кредитної історії: 1 – є кредити, оплата по яких прострочена; 2 – не було; 3 – є кредити у інших банках та погашаються своєчасно; 4 – є кредити у даному банку та погашаються своєчасно; 5 – немає відкритих кредитів, попередні сплачувались своєчасно; сума кредиту: 1000 – 100000 (грн.); строк кредитування: від 4 місяців до 5 років.

У якості вихідного параметру маємо лінгвістичну змінну, яка відповідає за ймовірність повернення кредиту та має такі терми: «дуже низька ймовірність

повернення», «низька ймовірність повернення», «середня ймовірність повернення», «висока ймовірність повернення», «дуже висока ймовірність повернення». Приклади реалізації функцій належності для змінних подані нижче (рис. 4.8, 4.9).

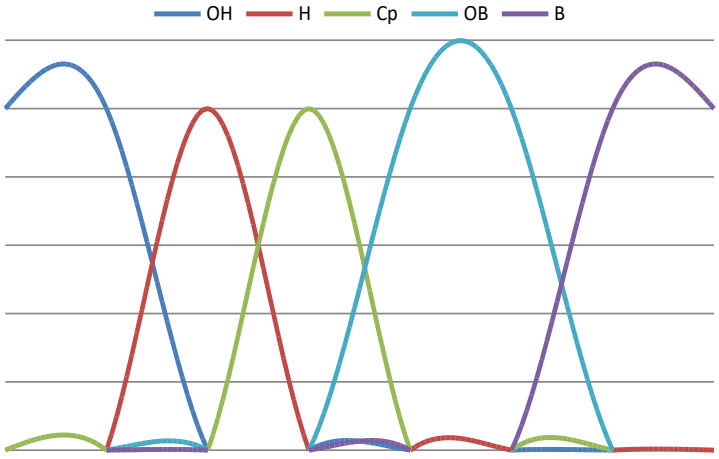

Рис.4.8– Функція належності змінної «Вік»

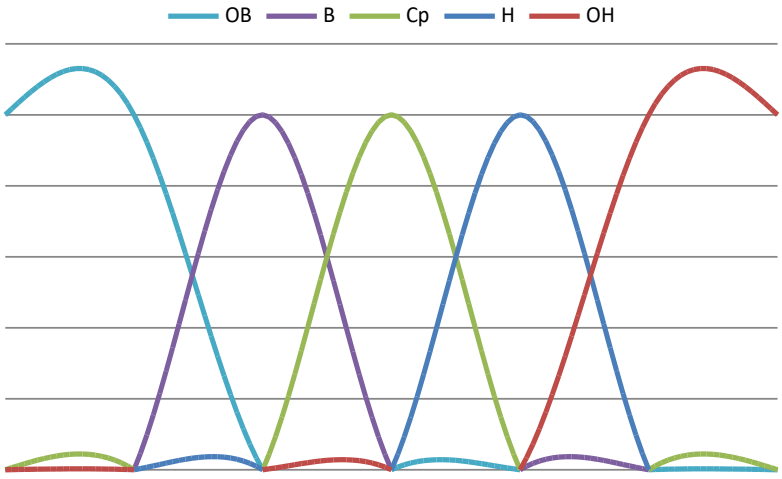

Рис.4.9 – Функція належності змінної «Спроможність погашення»

Змінна «Спроможністьпогашення» визначаєтьсятак:

$$Спроможність_{погашення} = \frac{Сума_{кредиту} Строк_{кредитування}}{Заробітна_{плата}}$$

Для запису правил, щовикористовувались у програмі, введемопозначення: «дуженизький» –ДН, «низький» – Н, «середній» – Ср, «високий» – В, «дужевисокий» – ДВ.

Аналіз результатів

За розглянутими моделями було виконано аналіз прогнозів перевірочних значень по кредитам (300 записів з 1600), які було отримано з використанням логістичної регресії, дерев рішень, мережі Байєса та за нечіткої логіки. Результати експериментів було порівняно з реальними значеннями [2]. Після проведення ROC-аналізу, де було встановлено пороги відсікання на певних рівнях, тобто, якщо ймовірність дефолту перевищує вказаний рівень, то клієнт вважається таким, що не поверне кредит, отримано таблицю результатів для логістичної регресії (табл. 4.9).

Таблиця 4.9

Загальна точність моделі та помилки I-го та II-го роду. Розглядаються різні рівні порогу відсікання, які отримані для логістичної регресії.

	Повернення кредиту (0)	Дефолт (1)
Cut-off = 0,1		
Повернення кредиту (0)	57	183
Дефолт (1)	0	60
Загальна точність моделі = 61,9 %		
Cut-off=0,15		
Повернення кредиту (0)	86	154
Дефолт (1)	1	59
Загальна точність моделі = 67 %		
Cut-off=0,2		
Повернення кредиту (0)	109	131
Дефолт (1)	3	57
Загальна точність моделі = 70,3 %		
Cut-off=0,25		
Повернення кредиту (0)	151	89
Дефолт (1)	6	54

Загальна точність моделі = 76,5 %		
Cut-off=0,3		
Повернення кредиту (0)	184	56
Дефолт (1)	11	49
Загальна точність моделі = 79,2 %		

Значення максимальної точності моделі на рівні 79,2 % було досягнуто при значенні порогу відсікання 0,3. В цьому випадку модель пропускає 11 значень дефолтів та відсіює 23,3% добросовісних позичальників. Значення площі під ROC-кривою дорівнює: $AUC = 0{,}775$, а індекс GINI відповідно дорівнює: $GINI = 2 \cdot AUC - 1 = 0{,}549$. На основі навчальної вибірки з 1300 клієнтів також побудована мережа Байєса. Для перевірочної вибірки (300 випадків) обчислено ймовірності дефолтів (табл. 4.10).

Таблиця 4.10

Загальна точність моделі та помилки I-го та II-го. Розглядаються різні рівні порогу відсікання, які отримані для мережі Байєса

	Повернення кредиту (0)	Прогноз: Дефолт (1)
Cut-off=0,1		
Повернення кредиту (0)	3	237
Дефолт (1)	0	60
Загальна точність моделі = 50,6 %		
Cut-off=0,15		
Повернення кредиту (0)	3	237
Дефолт (1)	0	60
Загальна точність моделі = 50,6 %		
Cut-off=0,2		
Повернення кредиту (0)	22	218
Дефолт (1)	0	60
Загальна точність моделі = 54,6 %		
Cut-off=0,25		

Повернення кредиту (0)	23	217
Дефолт (1)	0	60
Загальна точність моделі = 54,8 %		
Cut-off=0,3		
Повернення кредиту (0)	146	94
Дефолт (1)	3	57
Загальна точність моделі = 77,9 %		

Найбільша точність моделі в даному випадку досягається на рівні 77 % при встановленні порогу 0,3; при даному експерименті буде пропущено 3 значення дефолти та відкинуто 39 % добросовісних позичальників. Значення площі під кривою становить: $AUC = 0{,}879$, а індекс GINI становить: $GINI = 2 * AUC - 1 = 0{,}758$. Результати оцінювання кредитоспроможності для різних порогів відсікання при використання нечіткої логіки подані в таблиці 4.11.

Таблиця 4.11

Загальна точність моделі та помилки I-го та II-го роду. Розглядаються різні рівні порогу відсікання, які отримані для нечіткої логіки [2]

	Повернення кредиту (0)	Дефолт (1)
Cut-off=0,1		
Повернення кредиту (0)	42	198
Дефолт (1)	0	60
Загальна точність моделі = 58,75 %		
Cut-off=0,15		
Повернення кредиту (0)	88	152
Дефолт (1)	0	60
Загальна точність моделі = 68,3 %		
Cut-off=0,2		
Повернення кредиту (0)	121	119
Дефолт (1)	0	60

Загальна точність моделі = 75,3 %		
Cut-off=0,25		
Повернення кредиту (0)	161	79
Дефолт (1)	2	58
Загальна точність моделі = 81,9 %		
Cut-off=0,3		
Повернення кредиту (0)	173	67
Дефолт (1)	3	57
Загальна точність моделі = 83,5 %		

Як видно з отриманих результатів, нечітка модель показує кращі значення загальної точності для порогів відсікання 0,2 і 0,25, а також відсіює менше добросовісних позичальників порівняно з даними для мережі Байєса. Значення площі під кривою для моделі на основі нечіткої логіки: $AUC = 0{,}8875$, а індекс GINI відповідно: $GINI = 2 * AUC - 1 = 0{,}775$, що є кращим результатом.

Точність моделі та кількість помилок I-го та II-го роду залежать від порогу відсікання, який буде встановлений банком. Встановлення порогу відсікання визначає не лише процент клієнтів, які було відсіяно, а також нижню границю ймовірності повернення кредиту, тобто те значення, нижче якого клієнт вважається таким, що не поверне кредит. Значення дефолту в такому випадку буде 0,1 або 0,2 та буде є статистично незначними, а тому поріг відсікання доцільно встановлювати на рівні 0,25 – 0,3. За допомогою використання даного підходу на основі нечіткої логіки можна покращити якість моделі, а також її здатність розрізняти клієнтів, які повернуть або не повернуть кредити та зменшити кількість некоректно відсіяних клієнтів. У порівняльній таблиці 4.11 подано результати застосування використаних методів оцінювання кредитоспроможності.

Отримані результати обчислювальних експериментів свідчать, що найкращі результати дає нечітка логіка, значення результату дорівнює 0,835 або 83,5%, також

хороший результат дає модель у формі мережі Байєса та дорівнює 0,779 або 77,9%. Високі значення точності моделі отримані за логістичною регресією 0,792 або 79,2%. Наведені результати ще раз показують доцільність використання нечіткої логіки, а також логістичної регресії, дерев рішень та мереж Байєса при оцінюванні кредитоспроможності позичальників кредитів.

Таблиця 4.12

Порівняльна таблиця характеристик для різних моделей

Назва методу	Індекс GINI	Значення AUC	Точність моделі	Якість моделі
Бінарна логістична регресія	0,5491	0,7745	79,2 %	Прийнятна
Дерева рішень	0,548	0,774	71,25 %	Прийнятна
Мережа Байєса	0,758	0,879	77,9 %	Прийнятна
Нечітка логіка	0,775	0,8875	83,5 %	Висока

4.5. Аналіз результатів моделювання ринкових ризиків

На сьогодні зростання фінансових ризиків привертає все більшу кількість учасників, основна мета яких це отримання прибутку з використанням високоліквідних площадок. Зростання такої кількості учасників також прияє росту та розвитку великої кількості математичних моделей, які необхідні для опису даного процесу та даних. Зазвичай, традиційні лінійні моделі часових рядів не можуть коректно враховувати всі наявні характеристики, які можуть включати в себе фінансові дані та через це вимагають значного розширення структури. Було проведено безліч досліджень, які виявили цілі ряди особливостей часових рядів прибутковості фінансових активів і їх волатильності – відсутність автокореляції, довгострокова пам'ять, кластеризації волатильності, умовну гетероскедастичність, ефект «важеля» та інші. Формальною мірою волатильності виступає дисперсія або

середньоквадратичне відхилення, які також широко використовуються для оцінювання ризиків.

На сьогодні запропонована велика кількість моделей, що описують поведінку часових рядів. Одною з найпоширених моделей є Авторегресійна модель умовної гетероскедастичності (ARCH), узагальнена (GARCH). Точний характер фінансово-економічних процесів дуже важливий для багатьох проблем в макроекономіці і фінансах, наприклад незворотні інвестиції, ціни на опціони, структура процентних ставок по термінами і загальні динамічні співвідношення для цін активів.

Фінансовий ринок багатьох країн наразі нестабільний, тому реагування на зміни ринкових показників може бути не швидким та не завжди вчасним. Це має прямий вплив на те, що велика кількість моделей, які засновані на нормальному законі розподілу, не можна застосовувати для українського фінансового ринку так як вони будуть давати значні похибки у результатах при прогнозуванні зміни цін фінансових показників. Саме це і являється основною проблемою щодо практичного застосування класичних методів оцінювання VaR, і насамперед потребує використання більш сильних та гнучких підходів, технологій та статистичних методів.

Для виконання обчислювальних експериментів розроблено і практично реалізовано алгоритм триконтурної процедури адаптації, який призначено для аналізу статистичних даних, а також для побудови моделей гетероскедастичних процесів, який використовуються у прогнозуванні волатильності та оцінюванні ринкових ризиків на основі побудованих моделей. Також даний алгоритм використовується для аналізу якості отриманої прогнозованої моделі. Основною метою для створення триконтурної адаптації була робота з методами моделювання фінансово-економічних процесів, їх оцінювання та прогнозування. Також важливим було застосування підходу на реальних даних, та отримання відповідного прогнозу стосовно подальшої роботи та розвитку подій задля прийняття управлінських рішень. Процес аналізу представляє собою декілька етапів, що зображені на рисунку 4.10.

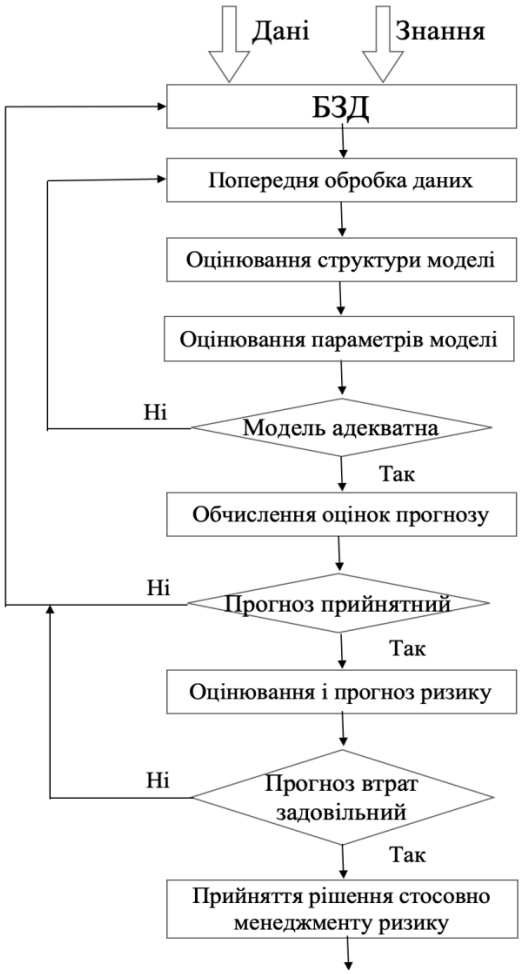

Рис. 4.10 Алгоритм трьоконтурної процедури адаптації

У наведеному експерименті часовий ряд був застосований для вивчення цін на золото в період між 2005-2006 роками. Статистичні характеристики, що показують побудовані моделі та якість прогнозів, наведені в таблиці.

В якості попередньої обробки даних використовувалась фільтрація. Початкові дані було пропущено через цифровий, гранулярний та оптимальній фільтри. Метою використання фільтрів є виконання згладжування даних (приглушення високочастотних компонент) і таким чином підготувати його до моделювання. Крім того, застосування оптимального ФК додатково забезпечує можливість вирішення наступних задач: оцінка компонентів вектора стану, що не піддаються вимірюванню; оцінка дисперсії (коваріації) для спостережень та вектора стану; і короткострокове прогнозування у випадку необхідності.

Використовувані дані відображають динаміку цін на акції золота разом з показниками технічного аналізу. Показники, обчислені на основі фактичної динаміки цін, створюють корисні дані для моделювання та прогнозування фінансових процесів. Для першого експерименту було взято початкові дані без фільтрації, на їх основі побудовано математичні моделі та прогноз.

Експеримент показав, що найкращою моделлю виявився AR (1) + тренд 4-го порядку. Він надає можливість на один крок вперед прогнозувати із середньою абсолютною процентною похибкою близько 3,19%, а коефіцієнт Тейла (Theil) - 0,024.

Таблиця 4.13

Показники якості моделей та прогнозу без застосування фільтра

Modeltype	Modelquality			Forecastquality			
	R^2	$\sum e^2(k)$	DW	MSE	MAE	MAPE	Theil
AR(1)	0,99	25644,67	2,15	49,82	41,356	8,37	0,046
AR(1,4)	0,99	25588,10	2,18	49,14	40,355	8,12	0,046
AR(1) + 1st ordertrend	0,99	25391,39	2,13	34,39	25,109	4,55	0,032
AR(1) + 4th ordertrend	0,99	25173,74	2,12	25,92	17,686	3,19	0,024

Коефіцієнт Тейла показує, що ця модель загалом корисна для короткострокового прогнозування.

Статистичні характеристики моделей та відповідні прогнози, обчислені за допомогою фільтра Калмана для згладжування даних, наведені в таблиці. В даному випадку оптимальний фільтр (із випадковою моделлю) відігравав позитивну роль, що підтримується відповідними статистичними параметрами кваліфікації.

Таблиця 4.14

Показники якості моделей та прогнозу із застосуванням фільтра Калмана

Modeltype	Modelquality			Forecastquality			
	R^2	$\sum e^2(k)$	DW	MSE	MAE	MAPE	Theil
AR(1)	0,99	23355,54	2,10	44,80	39,65	7,44	0,036
AR(1,4)	0,99	24132,15	2,08	46,45	38,34	6,78	0,034
AR(1) + 1stordertrend	0,99	23861,65	2,07	31,07	22,08	3,12	0,027
AR(1) + 4thordertrend	0,99	21887,54	2,05	21,24	13,13	2,55	0,017

Знову кращою моделлю виявився AR (1) + тренд 4-го порядку. Він надає можливість на один крок вперед прогнозувати з середньою абсолютною процентною помилкою близько 2,55%, а коефіцієнт Тейла: 0,017. Таким чином, у цьому випадку досягнуті результати є кращими, ніж у попередньому моделюванні та короткотерміновому прогнозуванні без застосування фільтра.

Статистичні характеристики моделей та відповідні прогнози, обчислені за допомогою гранулярного фільтра для згладжування даних, наведені в таблиці 4.15. В даному випадку фільтрації гранулярний фільтр відігравав позитивну роль, результати обчислювальних експериментів показали кращі результати ніж при фільтрації за допомогою фільтра Калмана та без фільтрації.

Таблиця 4.15

Показники якості моделей та прогнозу із застосуванням гранулярного фільтра

Model type	Model quality			Forecast quality			
	R^2	$\sum e^2(k)$	DW	MSE	MAE	MAPE	Theil
AR(1)	0,99	23312,32	2,09	44,78	39,45	7,18	0,035
AR(1,4)	0,99	24121,15	2,07	46,34	38,56	6,23	0,033

AR(1) + 1st ordertrend	0,99	23834,45	2,04	31,54	22,01	3,02	0,026
AR(1) + 4th ordertrend	0,99	20993,32	2,05	20,87	12,92	2,35	0,018

Кращою моделлю виявився AR (1) + тренд 4-го порядку. Середня абсолютна процентна помилка близько 2,35%, а коефіцієнт Тейла: 0,018.

Останнім підходом до фільтрації є експоненційне згладжування, результати якого показали, що як і у попередніх випадках, найкращий результат було виявлено у моделі AR (1) + тренд 4-го порядку (таблиця 4.16). Значення середньої абсолютної помилки у процентах складає 3,05% та значення коефіцієнт Тейла: 0,022.

Таблиця 4.16

Показники якості моделей та прогнозу із застосуванням експоненційного згладжування

Model type	Model quality			Forecast quality			
	R^2	$\sum e^2(k)$	DW	MSE	MAE	MAPE	Theil
AR(1)	0,99	24335,12	2,13	50,08	40,12	7,98	0,040
AR(1,4)	0,99	24453,1	2,12	47,15	39,56	7,68	0,042
AR(1) + 1st ordertrend	0,99	25061,08	2,10	34,07	23,26	3,55	0,033
AR(1) + 4th ordertrend	0,99	23881,14	2,07	24,1	16,58	3,05	0,022

Статистичний аналіз часових рядів, відібраних із застосуванням тесту Гольдфельда-Квандта, показав, що дані про ціни на золото створюють гетероскедастичний процес із залежною відмінністю у часі. Оскільки дисперсія є одним з ключових параметрів, який використовується в правилах здійснення торгових операцій, необхідно побудувати відповідні моделі прогнозування. Таблиця 4.17 містить статистичні характеристики побудованих моделей, а також якість короткотермінового прогнозу дисперсії. Для вирішення проблеми ми моделюємо GARCH разом з описом тенденції процесів, яка є досить складною (процес високого порядку). Моделі цього типу GARCH демонструють низьку якість короткотермінових прогнозів і цілком прийнятний EGARCH на крок вперед прогнозних властивостей.

Таблиця 4.17

Результати моделювання та прогнозування умовної дисперсії

Model type	Model quality			Forecast quality			
	R^2	$\sum e^2(k)$	DW	MSE	MAE	MAPE	Theil
GARCH(1,7)	0.99	153639	0.113	972.5	–	517.6	0.113
GARCH (1,10)	0.99	102139	0.174	458.7	–	211.3	0.081
GARCH (1,15)	0.99	80419	0.337	418.3	–	121.6	0.058
MLNSVM (3, 7)	0.99	61377	0.405	79.5	–	9.97	0.027
EGARCH (1, 7)	0.99	45184	0.429	67.8	–	8.74	0.023

Таким чином, найкращою побудованою моделлю був експоненційний GARCH (1,7). Досягнуте значення MAPE = 8,74% містить дуже хороший результат для прогнозування умовної дисперсії. Другою була запропонована модель MLNMSV (modified log-normal model of stochastic volatility), яка має два параметри порядку та MAPE = 9,97%. Подальше вдосконалення прогнозів було досягнуто із застосуванням схеми адаптації. Середнє поліпшення прогнозів було в межах від 0,5 до 1,5%, що виправдовує переваги запропонованого підходу. Поєднання прогнозів, створених за допомогою різних методик прогнозування, сприяло подальшому зменшенню середньої абсолютної похибки прогнозування приблизно на 0,5 – 0,8% у цьому конкретному випадку. Слід підкреслити, що аналіз гетероскедастичних процесів сьогодні дуже популярний завдяки безлічі інженерних, економічних та фінансових застосувань моделей та прогнозів, що ґрунтуються на них.

Проведемо оцінку ризикової вартості та очікуваного дефіциту за побудованими моделями волатильності за рекомендації Базельського комітету з банківського нагляду для довірчих інтервалів 95%, 97% та 99%. Результати порівняння та верифікації моделей для оцінювання ризику подано у таблиці 4.18. В дужках наведено загальну кількість перевищень (абсолютне значення).

Для визначення стандартного відхилення, яке є мірою волатильності при використанні дельта-нормального методу, повинно виконуватися припущення про

нормальний розподіл. Застосуємо дельта-нормальний метод для оцінки VaRбез перевірки історичних даних на «нормальність». У таблиці представлено результати прогнозування вартості золота для різних довірчих рівнів (95%, 97% та 99%).

Таблиця 4.18

Таблиця результатів проведення бек-тестування розрахунку VaR дельта-нормальним методом.

Тип моделі	Результати бек-тестування					
	95%		97%		99%	
	Кіл-ть перевищень	% вірних прогнозів	Кіл-ть перевищень	% вірних прогнозів	Кіл-ть перевищень	% вірних прогнозів
GARCH(1,7)	35	91,86%	36	90,65%	29	92,089%
GARCH (1,10)	48	85,13%	45	87,07%	40	88,96%
GARCH (1,15)	56	77,25%	51	78,98%	47	82,45%
MLNSVM (3, 7)	29	92,78%	27	93,13%	25	93,77%
EGARCH (1, 7)	62	79,01%	64	78,76%	58	78,92%

Для розрахунку кількості перевищень та проценту правильних прогнозів було застосовано VaR методу історичного моделювання.

На вхід моделі подаються ринкові значення ціни на золото. Для перевірки адекватності моделі використовуються рекомендації Базельського комітету з банківського нагляду для різних рівнів довіри (95%, 99% та 97%).

Таблиця 4.19

Таблиця результатів проведення бек-тестування розрахунку VaR методу історичного моделювання.

Тип моделі	Результати бек-тестування					
	95%		97%		99%	
	Кіл-ть перевищень	% вірних прогнозів	Кіл-ть перевищень	% вірних прогнозів	Кіл-ть перевищень	% вірних прогнозів
GARCH(1,7)	24	95,76%	18	96,62%	14	97,19%
GARCH (1,10)	32	89,22%	30	90,34%	26	91,15%
GARCH (1,15)	36	79,98%	51	80,08%	47	81,2%

MLNSVM (3, 7)	20	93,34%	27	94,67%	25	94,92%
EGARCH (1, 7)	55	81,12%	64	81,54%	58	83,15%

Точність за методом Монте-Карло майже не залежить від рівня довіри і на графіках чітко видно, що модель досить швидко пристосовується до змін на ринку. Для перевірки адекватності моделі використовуються рекомендації Базельського комітету з банківського нагляду для різних рівнів довіри (95%, 99% та 97%). Для кожної моделі підраховується кількість помилок прогнозів. Глибина ретроспективи для оцінки середнього і стандартного відхилення для побудови траєкторії руху курсів валют портфелю – аналогічна з попереднім експериментом. Результати верифікації моделі зведені у таблиці.

Таблиця 4.20

Таблиця результатів проведення бек-тестування розрахунку VaR за методом імітаційного моделювання Монте-Карло

Тип моделі	Результати бек-тестування					
	95%		97%		99%	
	Кіл-ть перевищень	% вірних прогнозів	Кіл-ть перевищень	% вірних прогнозів	Кіл-ть перевищень	% вірних прогнозів
GARCH(1,7)	16	96,12%	10	97,00%	7	97,92%
GARCH(1,10)	25	93,89%	17	95,14%	12	96,9%
GARCH (1,15)	28	91,01%	21	92,15%	16	95,7%
MLNSVM (3, 7)	13	97,39%	9	98,55%	3	99,04%
EGARCH (1, 7)	34	88,76%	27	91,13%	19	93,25%

За результатами, що наведені у таблицях, метод імітаційного моделювання Монте-Карло і метод історичного моделювання показують найкращі результати, прогнозні значення збитків покривають реальні у більшості випадків. Моделі є адекватними, проте більш точною є модель оцінки VaR за методом історичного моделювання [35, 36]. Дельта-нормальний метод є неадекватним для оцінки ризику, як показують результати прогнозування, і, що підтверджується

результатами проведення бек-тестування. Результати аналізу недоліків і переваг методів для оцінки VaRзведено у таблиці 4.21.

Таблиця 4.21

Порівняльний аналіз роботи різних методів для оцінювання ризику VaR.

Метод Критерії	Дельта - нормальний	Історичного моделювання	Метод імітаційного моделювання Монте-Карло
Оцінювання	Локальне	Повне	Повне
Врахування історичного розподілу	Як оцінка нормального розподілу	Аналогічно, тому, яке було у минулому	Повністю
Врахування «допустимої» волатильності	Можливе	Ні	Так
Припущення про нормальний розподіл доходностей	Так	Ні	Ні
Оцінка екстремальних подій	Погана	Погана	Можлива
Модельний ризик	Може бути значним	Допустимий	Високий
Об’єм ретроспектив	Середній	Дуже великий	Малий
Обчислювальна складність	Невисока	Висока	Дуже висока
Наглядність	Середня	Висока	Низька
Обчислювальні потужності	Низькі	Середні	Високі

Міра ризику VaRстала загальною мірою для оцінювання ринкових ризиків. Вона має ряд недоліків і переваг, які можна вважати як значними так і не значними, проте вона дає можливість оцінювати ризик уніфіковано для кожної країни і кожного банку.

Модель для оцінювання ризиків на основі дельта-нормального методу виявилася недостатньо адекватною через невиконання припущення про нормальний розподіл. Слід зазначити, що розподіл цін на золото на деяких періодах є близьким до нормального, тому модель оцінки VaR на цих періодах виявляється адекватною. Метод історичного моделювання показав задовільний результат лише за умов стабільної ситуації на ринку. Метод погано пристосовується до різких коливань на ринку і тому в нинішніх умовах не може використовуватися для поставленої задачі [36].

Найкращий результат оцінки можливих збитків показав метод імітаційного моделювання Монте-Карло, який гіпотетично враховує всі можливі зміни на ринку. Помилки у прогнозах можливих втрат трапляються лише за непередбачуваних різких змінах вартості, але модель на основі цього методу швидко пристосовується до змін на ринку.

4.6 Висновки до розділу

Початкові дані та атрибути, що впливають на кредитоспроможність позичальника можуть бути суперечливими або неповними. На кредитоспроможність позичальника, зазвичай, впливають фактори фінансового стану, до яких можна віднести власні кошти, ліквідні активи, та фінансову дисциплінованість. Протягом двох етапів, за якими складається оцінка кредитоспроможності необхідно враховувати та аналізувати діловий ризик; аналізувати фінансовий стан позичальника для якого враховуються фінансові коефіцієнти і грошові потоки. В таких випадках краще використовувати підхід, що складається із декількох методів аналізу даних та побудови прогнозів. Проведені експериментальні дослідження показали, що оцінка кредитоспроможності на

підставі комбінованого підходу дає кращі результати і дозволяє побудувати портрет позичальника та процес прийняття рішення.

Управління ринковими ризиками є проблемою, з якою стикаються всі учасники фінансового ринку. В умовах непередбачуваності, невизначеності і нестабільності ринку задача управління фінансовими ризиками повинна враховувати його проблеми і особливості із використанням методології VaR для оцінки ринкових ризиків. Базельський комітет з банківського нагляду визначає VaR як максимальне очікування збитків від прийняття ризику та як основний показник для оцінки ринкового ризику. Результати роботи дельта-нормального методу є менш задовільними через невиконання припущення про нормальний розподіл дохідностей коштів на метали. Нормальний розподіл виконується лише на деяких періодах при продажі і для цих періодів розрахунок можливих є адекватним. Метод історичного моделювання, який моделює можливі збитки лише за збитками в минулому, працює лише за умов стабільної ситуації на ринку. Серед переваг методу слід відзначити його простоту, точність оцінок і відсутність припущення про нормальність розподілу даних. Модель оцінки ризику VaR побудована за цим методом залишається адекватною навіть за кризової ситуації на ринку.

ВИСНОВКИ

У дисертаційному дослідженні розв’язана задача побудови адекватних математичних моделей і розробка ефективних методів розв’язання задач моделювання і прогнозування фінансових процесів та оцінювання ризиків можливих втрат шляхом використання методів інтелектуального аналізу даних, прикладної статистики і методів аналізу часових рядів.

За результатами дисертаційної роботи можна зробити наступні висновки:

1. Проведений аналіз існуючих математичних постановок задач, моделей та методів інтелектуального аналізу даних, в тому числі для кредитних та ринкових ризиків, що застосовуються і в інших сферах для розв’язання задач із ризиками та фінансовими даними.

2. Вперше розроблено метод оцінювання ринкового ризику на основі інтегрованого застосування ймовірнісної, оптимальної та цифрової фільтрації і регресійної моделі, який відрізняється високою якістю попередньої обробки даних і забезпечує підвищення якості оцінок прогнозів. Розроблений метод дозволив зберегти всі початкові дані та взаємозв'язки між ними. Проведені експерименти щодо оцінки ефективності запропонованих методів фільтрації показали, що дані методи дозволили покращити загальну точність початкових даних на 1.7%.

3. Вперше розроблено метод адаптації математичної моделі фінансового процесу до даних, який відрізняється застосуванням триконтурної процедури адаптації і забезпечує побудову моделі нелінійного процесу у вигляді лінійної та нелінійної компонент. Застосування даного підходу дозволило адаптувати процес моделювання нелінійного процесу, зменшити час аналізу та уникнути ручного перебору побудови великої кількості математичних моделей.

4. Удосконалено метод підвищення якості прогнозування кредитоспроможності позичальників який відрізняється комбінованим підходом до вибору регресорів та використанням альтернативних методів прогнозування що забезпечує оптимізацію вагових коефіцієнтів оцінок окремих прогнозів.

Розроблений метод дозволив уникнути ручного виділення ознак з початкових даних, що впливають на кредитоспроможність позичальника, а подавати на вхід всі доступні дані, не зменшуючи їх інформативності.

5. Удосконалено метод оцінювання кредитоспроможності позичальників з використанням адаптивної Байєсівської мережі, який відрізняється підвищеною адекватністю ймовірнісної моделі і забезпечує зменшення величини кредитного ризику. Експерименти щодо оцінювання кредитоспроможності запропонованими методами показали, що з використанням даних підходів було покращено загальну точність результатів даних на 2.5%.

СПИСОК ВИКОРИСТАНИХ ДЖЕРЕЛ

1. Бідюк П. І. Аналіз кредитоспроможності за допомогою методів інтелектуального аналізу даних / П. І. Бідюк, В. Г. Гуськова // Електронне моделювання. - 2019. - Т. 41, № 2. - С. 111-120.

2. Гуськова В. Г. Оцінювання кредитоспроможності позичальників кредитів методами інтелектуального аналізу даних / В. Г. Гуськова, П. І. Бідюк. // Міжнародній науково-технічний журнал «Системні дослідження та інформаційні технологі. – 2019. – №2. – С. 31–48.

3. Tymoshchuk O. L. A combined approach to modeling nonstationary heteroscedastic processes / O. L. Tymoshchuk, V. H. Huskova, P. I. Bidyuk. // Radio Electronics, Computer Science, Control. – 2019. – №2. – С. 80–89.

4. Гуськова В. Г. Розробка сценарного підходу на основі моделей інтелектуального аналізу даних / В. Г. Гуськова, П. І. Бідюк // Наукові праці Донецького національного технічного університету. Серія : Інформатика, кібернетика та обчислювальна техніка. - 2016. –№2. - С. 158-164.

5. Гуськова В. Г. Система підтримки прийняття рішень для прогнозування фінансових процесів на основі принципів системного аналізу / В. Г. Гуськова, Данилов В.Я., П. І. Бідюк, О.Л. Жиров // Міжнародній науково-технічний журнал «Системні дослідження та інформаційні технологі. – 2019. – №1. – С.20–36.

6. Гуськова В. Г. Аналіз кредитоспроможності позичальників кредитів за допомогою логістичної регресії. / В. Г. Гуськова, П. І. Бідюк // Наукові праці Донецького національного технічного університету. Серія : Інформатика, кібернетика та обчислювальна техніка. – 2017. –№2. - С. 54-60.

7. Zaychenko Yu.: Recognition of Objects on Optical Images in Medical Diagnostics Using Fuzzy Neural Network Neffclass / Yu. Zaychenko, V. Huskova // International Journal "Information Models and Analyses". 2016. – №5. pp. 13-22.

8. Бідюк П.І., Гуськова В.Г.: Застосування нечітких правил регресійного аналізу до фінансових даних, Institute for Modelling in Energy Engineering, NASc of Ukraine, NASc of Ukraine, September 12-14, 2018, Kyiv, Ukraine.

9. Huskova V.H., Bidyuk P.I.: Estimating financial risk using systemic approach, Проблеми інформатизації, тези доповідей шостої міжнародної науково-технічної конференції, 14 – 16 листопада 2018 року, Черкаси – Баку – Бельсько-Бяла – Харків.

10. Bidyuk P., Huskova V., Terentiev O.: Client solvency estimation using intellectual data analysis approach, The VIth International Conference «Advanced Information Systems and Technologies, AIST 2018», 16-18 May 2018, Sumy, Ukraine.

11. Zaychenko Yu., Huskova V.: Application of fuzzy neural network nefclass for recognition of medical images in diagnostics, System Analysis and Information Technologies; 18-th International Conference SAIT, 30 May - 02 June 2016, Kyiv, Ukraine.

12. Huskova V., Bidyuk P.: A Combined Approach to Modeling Heteroscedastic Processes and Financial Risk Estimation, Всеукраїнська науково-практична конференція комп’ютерна інженері яі кібербезпека: досягнення та інновації 27–29 листопада 2018 року, Кропивницький, Україна.

13. Гуськова В.Г., Бідюк П.І.: Побудова сценаріїв із використанням байєсівських методів, міжнародна науково-технічна конференція "моделювання і комп’ютерна графіка", 18–24 вересня, 2017 року, Покровськ, Україна.

14. Клапків М.С. Страхування фінансових ризиків: Монографія. – Тернопіль: Економічна думка, Карт-бланш, 2002. – 570 с.

15. Blank, I.A. (1999), Osnovy finansovogo menedzhmenta [Fundamentals of financial management], Vol. 2, Nika-Tsentr, Kyiv, Ukraine, 512 p.

16. Управління ризиками банків: монографія у 2 томах. Т. 2: Управління ринковими ризиками та ризиками системних характеристик / [А.О. Єпіфанов, Т.А. Васильєва, С.М. Козьменко та ін.] /За ред. Д-ра екон. наук, проф. А.О. Єпіфанова і д-ра екон. наук, проф. Т.А. Васильєвої. — Суми: ДВНЗ "УАБС НБУ", 2012. — 299 с.

17. Allen S. Financial risk management: A practitioner’s guide to managing market and credit risk / Allen S. – Hoboken, N.J.: John Wiley & Sons, Inc., 2003. – 567p.

18. Basel Committee on Banking Supervision (1996), "Amendment to the Capital Accord to Incorporate Market Risk", available at: http://www.bis.org/publ/bcbs24.pdf (Accessed 30 January 1996).

19. Внедрение нормативов Базеля II в банковский надзор Беларуси [Електронний ресурс] – Режим доступу : http://www.cbonds.info/by/rus/news/ index.phtml/params/id/379207

20. Basel Committee on Banking Supervision. Basel III: International framework for liquidity risk measurement, standards and monitoring. – Bank for International Settlements, December 2010. [Електронний ресурс]. – Режим доступу : http://www.bis.org/publ/bcbs188.pdf

21. R.A. Jarrow, S.M. Turnbull / Journal of Banking & Finance: North-Holland, 24 (2000) pp. 271-299.

22. Elena Medova, Robert Smith, 2005. "A framework to measure integrated risk," Quantitative Finance, Taylor & Francis Journals, vol. 5(1), pp. 105-121.

23. Theodore M. Barnhill and William F. Maxwell Journal of Banking & Finance, 2002, vol. 26, issue 2-3, pp. 347-374.

24. Xeni Kristine Dimakos and Kjersti Aas, Integrated risk modelling in Statistical Modelling 4(4): pp. 265-277 – December 2004.

25. Тен В. В. Проблемы анализа кредитоспособности заемщика / В. В. Тен //Банковское дело. – 2006. – № 3.

26. Crouhy M., Mark R., 1998. “A Comparative Analysis of Current Credit Risk Models,” Manuscript, Conference on Credit Risk Modelling and Regulatory Implications.

27. Federal Reserve System Task Force on Internal Credit Risk Models, 1998. “Credit Risk Models at Major U.S. Banking Institutions: Current State of the Art and Implications for Assessments of Capital Adequacy.” Manuscript, Board of Governors of the Federal Reserve System.

28. Credit Suisse Financial Products, 1997. CreditRisk+: A Credit Risk Management Framework. (http://www.csfp.co.uk/csfpfod/html/csfp_10.htm).

29. Berkowitz, J., 1999. “Evaluating the Forecasts of Risk Models,” Manuscript, Trading Risk Analysis Group, Federal Reserve Board of Governors.

30. Бідюк П.І., Романенко В.Д., Тимощук О.Л. Аналіз часових рядів: навчальний посібник Київ: НТУУ «КПІ», 2013. 600 с.

31. Галасюк В. В. Методика оцінки кредитоспроможності позичальників // Вісник НБУ. – 2009. – № 2. – С. 39.

32. Бідюк П. І. Моделі оцінки ризиків кредитування фізичних осіб / П. І. Бідюк, Є. О. Матрос // Кібернетика та обчислювальна техніка. – 2007. – №153. – С. 87–95.

33. Морозов В.С. Регулювання банківського сектору Німеччини // Інформаційно-аналітична довідка. – Торговельно-економічна місія у складі Посольства України у ФРН. – Берлін, 2007. – 12 с.

34. Siconolfi M., Raghavan A., PacelleStaff M. All Bets Are Off: How the Salesmanship And Brainpower Failed At Long-Term Capital (англ.) // Wall Street Journal : newspaper. — 1998. — 16 November.

35. Lopez, J.A., 1999b. “Methods for Evaluating Value-at-Risk Estimates,” Federal Reserve Bank of San Francisco Economic Review, forthcoming.

36. Lopez, J.A., 1999a. “Regulatory Evaluation of Value-at-Risk Models,” Journal of Risk, forthcoming.

37. Міжнародні стандарти фінансової звітності (МСФЗ, МСФЗ для МСП, включаючи МСБО та тлумачення КТМФЗ, ПКТ) https://zakon.rada.gov.ua/laws/show/929_010

38. Paper 2751-2018 SAS® Credit Scoring for Banking – An Integrated Solution from Data Capture to Insight Ewa Nybakk, Capgemini Norway

39. Building Powerful, Predictive Scorecards An overview of Scorecard module for FICO® Model Builder

40. Маккинли У. Python и анализ данных. — Перевод с английского. — М.: ДМК Пресс, 2015. — 482 с. — ISBN 978-5-9706-0315-4.

41. И. А. Хахаев. Практикум по алгоритмизации и программированию на Python. Учебник. — М.: Альт Линукс, 2010. — 126 с. — (Библиотека ALT Linux). — ISBN 978-5-905167-02-7.

42. An Introduction to Statistical Learning / with Applications in R Authors:

James, G., Witten, D., Hastie, T., Tibshirani, R.

43. Bermingham, M.L. et al. Application of high-dimensional feature selection: evaluation for genomic prediction in man. Sci. Rep. 5, 10312; doi: 10.1038/srep10312 (2015).

44. Isabelle Guyon, André Elisseeff. An Introduction to Variable and Feature Selection // JMLR. — 2003. — T. 3.

45. Meuwissen, T. H., Hayes, B. J. & Goddard, M. E. Prediction of total genetic value using genome-wide dense marker maps. Genetics 157, 1819–1829 (2001).

46. Urbanowicz RJ, Meeker M, La Cava W, Olson RS, Moore JH. Relief-based feature selection: Introduction and review. J Biomed Inform. 2018;85:189-203. doi:10.1016/j.jbi.2018.07.014

47. Forman, 2003, c. 1289–1305.

48. Gengjian Xue, Li Song, Jun Sun. Foreground estimation based on linear regression model with fused sparsity on outliers, IEEE transactions on circuits and systems for video technology. 2013: pp. 1346-1357.

49. A Rakotomamonjy, FR Bach, S Canu, Y Grandvalet, SimpleMKL, Journal of Machine Learning Research 9 (Nov), pp. 2491-2521.

50. Орлов А.І. Прикладна статистика/А.І.Орлов. М.: Іспит, 2004. 284 с.

51. (А. Романовского, Б.Фішера та ін)

52. Айвазян С.А. Теорія ймовірностей та прикладна статистика / С.А.Айвазян. М.: Юніті, 2001. – 656 с.

53. Мостеллер Ф. П'ятдесят цікавих імовірнісних задач з рішеннями/ Ф. Мостеллер. М.: Наука, 1975. 111 с.

54. Чистяков В.П. Курс теорії ймовірностей/В.П.Чистяков. М.:Наука, 1982. 256 с.

55. H. Xiong and X. Chen, “Kernel-based distance metric learning for microarray data classification,” BMC Bioinformatics, 7:299, 2006.

56. J. Weston, A. Elisseeff, B. Scholkopf, and M. Tipping, “Use of the zero-norm with linear models and kernel methods,” Journal of Machine Learning Research,

special Issue on variable and Feature Selection 3, pp.1439-1461, 2003

57. I. Guyon, J. Weston, S. Barnhill, and V. Vapnik, “Gene selection for cancer classification using support vector machines,” Machining Learning, vol. 46, no. 1-3, pp. 389-422, 2002.

58. Ларичев О.И. Качественные методы принятия решений. Вербальный анализ решений / О.И. Ларичев, Е.М. Мошкович. – М., 1996.

59. Терентьев А. Н. Сравнение методов интеллектуального анализа данных при оценивании кредитоспособности физических лиц / [А.Н. Терентьев, П.И. Бидюк, Миронова А.В., Мединн Н.Ю.] // Проблемы управления и информатики. – К.: ИКИ НАНУ-НКАУ, 2009. – № 5. – С. 141–149.

60. Бідюк П.І. Основні етапи побудови і приклади застосування мереж Байєса/ Бідюк П.І., Кузнєцова Н.В.// Системні дослідження та інформаційні технології. – 2007. – № 4. – С.26–39.

61. Heckerman D. Bayesian Networks for Data Mining// Data Mining and Knowledge Discovery. – 1997. – № 1. – P. 79 – 119.

62. Кузнєцова Н.В. Системний підхід до аналізу кредитних ризиків з використанням мереж Байєса / Н.В. Кузнєцова, П.І. Бідюк // Наукові вісті НТУУ «КПІ». – 2008. – № 3. – С. 11 – 24.

63. Недосекин А. Методологические основы моделирования финансовой деятельности с использование нечетко множественных описаний / Недосекин А. – СПб: 2003. – 280 с.

64. Зайченко Ю.П., Оценка кредитных банковских рисков с использованием нечеткой логики// Системні дослідження та інформаційні технології. – 2010. –No2. – с. 37-54.

65. Шовгун Н.В. (Шаповал Н. В.) Аналіз кредитоспроможності позичальника за допомогою методів з нечіткою логікою/ Н.В. Шовгун // Вісник НТУУ «КПІ». Інформатика, управління та обчислювальна техніка: Зб.: наук. пр. – К.: Век+,- 2012. - №55. – С.169-173

66. Тест Голдфелда-Квандта . Голдфелдом і Р. Квандтом у 1956 р

67. Kim, E.H.; Morse, A.; Zingales, L. (2006). «What Has Mattered to Economics since 1970». Journal of Economic Perspectives 20 (4): 189—202. doi:10.1257/jep.20.4.189

68. Harvey A., Shepard N. Estimation of an asymmetric stochastic volatility model for asset returns // Journal of Business and Economic Statistics. – 1996. – Vol. 14, No.4. – Pp. 429- 434.:

69. Kloeden, P.E. & Platen, E. (1992). Numerical Solution of Stochastic Differential Equations. Springer, Berlin. ISBN 3-540-54062-8.

70. Broze L., Scaillet O., Zako J. Quasi-indirect inference for diffusion processes / Broze L., Scaillet O., Zako J. // Econometric Theory.– 1998. – Vol. 14, № 02. – Pp. 161-186.

71. Affet-Sahalia Y. Transition Densities for Interest Rate and Other Nonlinear Diffusions // The Journal of Finance. – 1999. – Vol. 54, No. 4.– Pp. 1361–1395.

72. Бідюк П.І., Коновалюк М.М. Прогнозування волатильності валютного ринку за нелінійними моделями // Вісник Національного університету «Львівська політехніка», № 719. –2011р., с. 154 – 163.

73. Hansen L. Large Sample Properties of Generalized Method of Moments Estimators / Hansen L. // Econometrica. – 1982. – Vol. 50, № 4. – Pp. 1029–1054.

74.P. Brown R.G. Smoothing forecasting and prediction of discrete time series. - N.Y., 1963.

75. Brown R.G., Meyer R.F. The fundamental theorum of exponential smoothing. Oper. Res. - 1961. - Vol.9. -№ 5.

76. Хемминг Р.В. Цифровые фильтры. – М.: Недра, 1987. – 221 с.

ДОДАТОК А
ДОКУМЕНТИ, ЩО ПІДТВЕРЖУЮТЬ ВПРОВАДЖЕННЯ РЕЗУЛЬТАТІВ ДОСЛІДЖЕНЬ